

Timely lessons in investment from a wealthy 'wonk'

President Bill Clinton's budget proposal contains an important message for governments that are tempted to undervalue basic science, its crucible, the research university, and the importance of impartial peer review.

Why, after five years of lukewarm support for basic scientific research, has the Clinton administration just proposed a generous budget which will invest an additional \$2 billion next year in the National Institutes of Health (NIH), the Department of Energy and the National Science Foundation (NSF; see pages 521–522)? One possibility is that the scientific community's own fightback, orchestrated by scientific societies after the 1995 threat of serious cuts in research funding, was largely responsible for the president's change of heart. Some have pointed out that President Bill Clinton was in danger of being outflanked on the issue by Republicans in the Congress, some of whom have called for substantial extra research funding. And the effectiveness and popularity of Harold Varmus, the director of the NIH, is noteworthy — although that hardly explains the good fortune of the other science agencies.

But a close examination of Clinton's increasingly vocal statements on science policy issues over the past two years points towards a more significant explanation of science's newly improved budget prospects. The president's speeches, and now his actions, suggest that Clinton, perhaps the most enthusiastic policy wonk ever to occupy the White House, has done to science what he likes to do best: hammered at the available political and economic data — and concluded that investment in research is a winner on both counts.

Priority

When Clinton and his vice-president Al Gore came to power in 1993, their priority in this arena was technology, not science. Their main initiative was the Advanced Technology Program at the Department of Commerce, which was supposed to provide \$1 billion a year to help corporations to exploit the country's science base, already felt to be strong. "Technology" came up repeatedly in speeches by both men and "science" not at all, and the inference was clear: science could look after itself, and it was the role of government to bridge the alleged gap between scientific and industrial innovation.

Clinton's great plans for technology support were frustrated by the Congress, which argued that they represent inappropriate intervention. But, notwithstanding the alleged gap, the economy of the United States powered ahead. At the same time, the German or Japanese economic models, with their heavy reliance on direct or indirect national support for technology, fell into disrepute. The world's economists turned to the United States: they saw flexibility in labour markets and sundry other intangibles. They also saw booming information technology and biotechnology industries, closely tied to top-class research universities. So did Clinton.

As governor of Arkansas (and as a Rhodes scholar at Oxford) Clinton had scant prior knowledge of the modern research university. But travelling around the United States as president, he could hardly miss these establishments. Although most of their money comes from government, their research activities are far more intimately linked to private-sector wealth generation than any government department or laboratory. From San Diego and Tucson to

Philadelphia and Pittsburgh, such institutions are now the largest employers in many US cities, as well as the hub of their cultural, intellectual and economic life. As Clinton toured the country during his 1996 campaign, it is little wonder that science and research began to permeate speeches in which he promised to build "a bridge to the twenty-first century" (see *Nature* 383, 566; 1996).

A parallel awakening stirred Republicans in the Congress. Last year, a bipartisan group of senators led by Phil Gramm (Republican, Texas) and Joe Lieberman (Democrat, Connecticut) called for a doubling of funding in research and development, while Newt Gingrich (Republican, Georgia), Speaker of the House of Representatives, identified biomedical research as one of three areas on which the government should spend more money (the military and new roads being the other two).

Investment

Real impetus for such investment has been slow to materialize in the Congress. A brief action plan issued last week by the Senate Republican leadership found room for a mischievous proposal to foist the name of Ronald Reagan on Washington National Airport, but not for the Gramm–Lieberman bill. But over the Christmas period, Clinton's conversion became complete. Revised budgetary projections allowed the administration to plan for new investments in science without creating a budget deficit. Clinton's State of the Union address last week was punctuated with references to research and culminated in a eulogy to science's ability to shape the new millennium.

Political rhetoric about the importance of science tends to be ignored by a press that is simultaneously bored by and suspicious of it. But the rhetoric, and the soaring approval ratings that Clinton enjoyed after the State of the Union address, should interest all denizens of science policy. It seems that the conventional wisdom — that the public is not interested in science; that there are no votes in it; that it carries no political capital — has recently been turned on its head. This year in the United States, the Congress and the administration are in a race to win credit for spending more money on science.

Of course, they have extra money — other people's money — to spend. They haven't had to do things the hard way, for example by clawing back some of the \$18 billion that the United States spends each year on a grossly inefficient network of government laboratories that neither Clinton nor the Congress has the guts to reform.

Indeed, few countries have the budgetary freedom to follow the United States' lead. But there are a couple of things from which they can learn, as they observe this latest expansion of the US research system. The first is the success of a flexible, multifaceted university system, in which public and private establishments compete freely and fiercely for money and students (see *Nature* 391, 8; 1998). The second is a set of research agencies such as NIH and NSF, both with an enviable record of fairness in awarding grants to the universities. Clinton is showing that money spent on science, if distributed wisely and fairly, can win public approval. The principle is universal. □



French ministry reopens inquiry into conduct of INSERM unit

[PARIS] France's ministry of national education, research and technology is to reopen an inquiry into the activities of a laboratory of INSERM, the national biomedical research agency, following criticisms by the university at which the laboratory is located — as well as some present and former staff — of the way in which the previous government handled earlier evaluations and investigations.

Controversy centres on claims by the INSERM Laboratory of Nutrition, Lipoprotein Metabolism and Atherosclerosis, based at the University of Rennes 1, to have identified and now cloned a 'lipolysis stimulated receptor', a molecule involved in fat degradation. "As there is a new government we felt we needed to have a fresh look at the matter," says Jacqueline Godet, director of life sciences at the ministry.

INSERM and the French biotechnology company Genset have jointly applied for a patent on the work — which may have implications for obesity research — as part of plans costing FF60 million (US\$10 million) to build a laboratory for molecular physiology and pathology at a new private university at Ker-Lann near Rennes. The project is being partly financed by local government.

The allegations have come mainly from a



group of more than 20 'whistleblowers' who work or have worked in the laboratory. The group refuse to reveal their identities publicly, although they have done so in confidence to recent inquiries. But Bernard Bihain, the director of the laboratory — and the principal target of the allegations — argues that this is unfair, and that it leaves him open to 'score settling' stemming from personal conflicts.

The only relatively detailed exposition of the scientific aspects of the affair available at

present is a report commissioned by the university from Michel Philippe, a researcher at the Centre National de la Recherche Scientifique laboratory of biology and genetics of development at the University of Rennes 1. The report, submitted in December to Jacques Lenfant, the president of the university, describes difficulties in "reconciling" results concerning the molecular weights of candidate clones of the receptor and its antibody reactivity.

Analysis of the allegations is complicated by the fact that the University of Rennes 1, which has demanded a broader investigation of the laboratory, is in dispute with INSERM over the terms of the industrial joint venture. Lenfant claims the university was not consulted on the initiative and was omitted from the patent applied for jointly by INSERM and Genset. INSERM "failed to respect" its contract with the university, he says.

Bihain claims that he is caught in a conflict between the university and INSERM over the patent. He denies any wrongdoing, and criticizes the inquiries for not allowing him due process, saying "I have never been given the opportunity to clear my name." He also argues that, even if exonerated, suspicion will inevitably be cast upon his work.

But Lenfant claims the importance of the industrial venture, under which Genset has worldwide rights to licences on any patents, with INSERM sharing royalties, may have influenced past investigations. The deal, announced last February, was described then by Pascal Brandys, Genset's chief executive officer, as "instrumental in expediting our massive search for disease susceptibility genes." INSERM's director general, Claude Griscelli, called it an example of the "policy of industrial partnerships that INSERM wishes to reinforce".

Lenfant claims that evidence of pressure to protect these interests comes from the circumstances surrounding a vote in May by the INSERM committee responsible for the routine evaluation of Bihain's laboratory, which narrowly approved — by 13 votes out of 25 — renewal of the laboratory.

Before the vote Griscelli emphasized to the meeting the importance to INSERM of pursuing the industrial partnership. Critics say that this constituted a form of pressure on the committee. This was "abnormal", says Lenfant, arguing that it is "astonishing that a director general should intervene at this level".

Jean-Claude Stoclet, chairman of the commission, denies that it was improper for Griscelli to say that the agreement exempli-

Spain blocks new Framework programme

[MUNICH] The smooth continuation of the European Union (EU)'s five-year Framework programmes of research is being held hostage for the first time to broader political considerations. At a meeting of representatives of member states in Brussels last week, Spain — supported by Portugal — refused to vote on the proposals for the fifth Framework programme (FP5), due to start next year.

Spain is trying to use its veto power to force member states to continue to give it large amounts of structural funds — subsidies given to poorer EU countries and regions. Because the Framework programme must be approved unanimously by the Council of Ministers, which represents the member states, the result of Spain's action could be a gap in funding for researchers, as the fourth Framework programme (FP4) ends next December.

The EU has been required to rethink the way in which it distributes its structural funds because of its planned enlargement within the next decade to include some eastern European countries. Last summer, the European Commission published

Agenda 2000, an informal communication covering the period 2000–06, proposing that, to make structural funds available to the newcomers before joining, spending on current members should not increase in nominal terms.

This means considerable real-term losses in income for major beneficiaries such as Spain, which is depending on these funds to maintain its economic criteria for joining full European monetary union. The threat to EU-funded research comes from the fact that Agenda 2000 will not be debated before the end of this year.

EU research ministers have their last chance next week to approve a common position, if FP5 is to start on schedule. Attempts are being made to persuade Spain to drop its blocking action. Its toughest opponent will be Germany, which strongly supports the move to divert structural funds to help future EU members that it borders, such as Poland and the Czech Republic. Germany opposes an increase in the overall budget to allow current beneficiaries to enjoy the level of subsidies to which they have become accustomed.

Alison Abbott

fied INSERM's industrial policy. But he accepts that the importance of the industrial venture constituted indirect pressure: "If we had given a negative opinion, we would have been disowning our own institution."

Bihain asserts that it was around this time that he discovered the university was investigating him "without his knowledge", and that he subsequently "telephoned the ministry to seek independent arbitration". In October he filed charges of defamation against Lenfant, although he has now dropped proceedings, understanding the affair to be closed.

On 15 September, Bernard Bigot, then director general of the research ministry, set up a commission of inquiry, chaired by Pierre Corvol of the Collège de France in Paris, which received testimony from more than 20 of the whistleblowers. It submitted its report to the ministry on 28 October.

But the report was never made public, and according to Lenfant the university has itself been unable to obtain a copy, despite a written request to Claude Allègre, the science minister, which he says has gone unanswered.

Jean Rey, an adviser to Allègre, says "we have not excluded [scientific fraud], but we have not established it."

Bihain says he has not seen the Corvol report, but he understands its main conclusions are that additional experiments to verify the identity of the receptor would have been useful. According to Bihain, it also shows that one gel lacked proper molecular weight standards, but he claims that this involved a flawed experiment by a young researcher that was never published.

In written testimony to the Corvol com-

mission, Daniel Renou, the INSERM representative in the region until last August, says Bihain treated those who worked for him harshly. "The institute cannot remain deaf to the distress provoked," he wrote.

Renou also submitted evidence of what he claimed were violations of health and safety regulations in the laboratory, and criticized what he describes as Bihain's lack of cooperation on these issues with the authorities. Corvol last week declined to comment on the report, saying only that the affair was "complex".

Controversy also surrounds the fact that the Corvol report was not made available to a meeting of the scientific board of INSERM, at which Bihain's unit was again evaluated, a week after the report was submitted to the ministry. The commission gave a positive decision on the basis of Bihain's presentation, including the fact that a patent had been applied for on his research, and that the results on which this was based had been submitted to a number of scientific journals. But it ruled that the laboratory should be evaluated in two years' time, rather than in four as is usual.

Rose Katz, who chairs the board, says she requested a copy of the Corvol report from Griscelli, but he replied that he did not have one. Katz, who says it would have been "preferable" for the board to have seen the report before ruling on Bihain's laboratory, adds that she has still not been able to see the report. It would have been "more healthy" for this to have been made public, she says.

To complicate matters further, Bigot then carried out his own inquiry, visiting the Rennes laboratory on 28 November and telephoning current and former staff. This

second inquiry was needed to "ensure the veracity of the testimonies," says Rey.

Bigot's report, submitted to the ministry on 10 December, cleared the laboratory of misconduct. A summary concludes that "no formal element" refuted the laboratory's research, which, alluding to the industrial interest, Bigot describes as being "on a promising subject that has recently known considerable developments, and which it would be unacceptable to see being abandoned". Nevertheless the report recommends that Bihain "be invited to pay greater attention to his behaviour vis-à-vis his collaborators". (Since 1993, more than 40 staff have left the laboratory or been fired.)

In a letter on 21 January to Daniel Nahon, the new director general of the science ministry, the whistleblowers call for the case to reopened, arguing that the Bigot report fails to take their allegations seriously. The letter — again anonymous, although the signatories' names have been deposited in confidence with Philippe — points out that the Bigot report cites only one paragraph from the Corvol report, stating that "the commission attracts the attention of the minister to the fact that it was convened, by a large number of members and former members of the INSERM unit 391, to consider serious accusations concerning the behaviour of the director of this unit and its scientific activity."

The Philippe report states that some of the laboratory's work poses "many problems". It concludes that "an in-depth analysis carried out by international experts in the field is absolutely necessary."

Declan Butler & Olivier De Gandt

New Hubble institute chief seeks broader focus and more work

[WASHINGTON] The institute that manages the science programme for the Hubble Space Telescope (HST) should expand its focus beyond that instrument, according to the director chosen to lead the institute into the next century.

Steven Beckwith, who is to take over the helm at the Space Telescope Science Institute (STScI) in Baltimore, Maryland, next September, says one of his main objectives will be bringing in new work to the institute. "HST is the only thing that's done by the institute, and I personally think that's too little for an institute with the kind of resources that are there," says Beckwith, who is at present managing director of the Max Planck Institut für Astronomie in Heidelberg, Germany.

Beckwith, whose appointment was announced by the Association of Universities for Research in Astronomy last week, succeeds Robert Williams as the institute's third director. A native of

Wisconsin, Beckwith was a member of the astronomy faculty at Cornell University before joining Max-Planck, which he helped turn into a leading astronomy centre. He is an expert on star and planet formation, and has been a principal user of Europe's Infrared Space Observatory.

The most obvious new line of work would be to manage the scientific programme for the planned Next Generation Space Telescope (NGST), an infrared-optimized space observatory that the US National Aeronautics and Space Administration plans to launch as a successor to Hubble around 2007 (see *Nature* 389, 651; 1997). The STScI is generally thought to be the leading contender for operating the NGST. But "even that by itself is probably not

enough", says Beckwith.

The institute, he says, could be a leader in providing software products, such as those used for data reduction and telescope scheduling, to the general astronomy community. Programs written for the space telescope could be used just as well by other space- and ground-based instruments.

Although some sharing is already going on, he says, the astronomy community has not reached the level of collaboration that he sees among particle physicists. "We don't share as much as we could. It should be possible to save money overall for the field, or at least be able to use the resources we've got to do more than we are, by not duplicating efforts."

This sharing of resources has political problems, he admits, and will happen only if large observatories and astronomy institutions sit down together to "establish a dialogue among the principal actors." Beckwith would like the institute to play a lead role in that dialogue. **Tony Reichhardt**



Beckwith: keen for more collaboration.

NSF returns to favour with biggest increase

Christmas has come late for the National Science Foundation (NSF), for which the Clinton administration is proposing a budget increase of 10 per cent, to \$3.8 billion, next year (see right) — more than any other major government science agency. A slightly lower percentage is allocated to the agency's education activities, so NSF's research directorates will receive, if the Congress implements the proposal, an unprecedented 12 per cent increase.

Neal Lane, director of NSF, says he is uncertain what led the administration finally to accept the case — long put by advocates of the agency — that its wide portfolio of science support deserves to benefit from the groundswell of public support for more biomedical research. "I don't think I could pick out what made the difference. But the strong and consistent support of Harold Varmus [director of NIH] has been very important."

The agency plans to use the money to increase the size and duration of its grants, while holding the success rate of grant applications steady at the current rate of 33 per cent. "We've been trying to push in this direction for a long time," says Lane. "But recently we'd lost a little ground." The agency plans to increase the average annual value of a grant from \$80,000 in 1997 to \$90,000 in 1999, and to stretch its average duration from 2.3 to 2.7 years.

In keeping with the administration's commitment to research on faster Internet links, the Computer and Information Science and Engineering (CISE) directorate fares best in the proposal, with an increase in funding of 16.5 per cent to \$330 million. The other research directorates get about 11 per cent rises each, and the new Plant Genome Initiative is funded at this year's level of \$40 million. **Colin Macilwain**

Biomedical research shines in Clinton budget bonanza

[WASHINGTON] President Bill Clinton this week promised record increases in funding for the National Institutes of Health (NIH) and the National Science Foundation (NSF) in a budget that delighted science advocates and confounded widespread expectations of tough times ahead for US researchers.

The Clinton administration is proposing increases of 8.4 per cent, or \$1.15 billion, at NIH for the 1999 financial year, which begins on 1 October. An unprecedented 12 per cent increase would be made in research funding at the NSF, which supports most non-biomedical research at US universities. Other agencies with strong basic research interests also fared well, as the administration dropped the emphasis on technology support that had characterized its previous five budgets.

Clinton had promised "the greatest increases in history" for NIH and NSF during his State of the Union address on 26 January, which was watched by a record television audience, largely because of recent accounts of his alleged marital infidelity. Harold Varmus, director of the NIH, was guest of honour at the address, and strategically seated next to the president's wife Hillary.

New projections published by the administration would increase the NIH budget by half to more than \$20 billion by 2003. Other non-defence research programmes would also prosper thanks to a '21st Century Research Fund', worth \$171 billion over five years, announced by Clinton.

According to White House officials, the fund has no special legal status. But \$25 billion is expected to come from new taxes on cigarettes which the Congress may enact as part of a 'tobacco agreement' later this year.

Announcing the science and technology budget proposal, Vice-President Al Gore described the five-year plan as "the largest commitment to key civilian research in the



Good times ahead: NIH director Harold Varmus and Hillary Clinton applaud her husband.

history of the United States". He added that "more and more we find that science and society are interconnected", and predicted that science would shortly cross "important cross-roads", including the complete sequencing of the human genome and the creation of "the first preventative cancer medicines".

Varmus described the increase for the NIH as being "of historic proportions". He added: "It is double what any president has ever asked for, and more than the Congress has ever appropriated."

Neal Lane, director of the NSF, said that the increase at his agency "set the stage for a new century of progress in knowledge and discovery. It is a great time to be a scientist, and even a good time to be a science bureaucrat."

The budget proposal would increase total

NASA cuts Earth observation, boosts 'gravitational biology'

Space agency chief Daniel Goldin, who has promised the White House that he can do more with less, will not be sharing in the federal research spending spree. NASA's 1999 budget request drops slightly from this year's approved level, to \$13.46 billion.

Nevertheless, Goldin says, "for what we have on our plate today, we have adequate resources", and all major initiatives are fully funded. Space science — which includes astronomy, solar-terrestrial studies, and planetary exploration — remains about level with last year's allocation, at \$2 billion.

The Discovery low-cost planetary programme, which has added two missions to its launch queue, gets a 65 per cent boost. Major development efforts, such as the Space Infrared Telescope Facility (SIRTF), the Origins programme and Mars exploration, stay on track, and the agency will begin work on a Europa orbiter to launch in 2003 as its first low-cost outer planets mission.

The Earth Observing System (EOS) of satellites takes a cut from \$704 million to \$659 million. But NASA will increase spending on smaller, more focused 'Earth probes'. While funding for microgravity

research stays flat, gravitational biology and biomedical research each increase by about 25 per cent, and funding for space station research equipment jumps dramatically, from \$221 million to \$374 million.

One possible setback for space scientists has emerged. Last year, Congress ordered NASA not to borrow money from other parts of the agency to pay for an overrun in space station funding. Goldin now says that the agency wants to transfer \$50 million in unspent funds from each of the Earth science and space science accounts to help solve the station's problem. **Tony Reichhardt**

basic research paid for by the government by 8 per cent, or \$1.2 billion, to \$17 billion, and applied research by 5 per cent, or \$850 million, to \$16.5 billion. Total federally funded research and development would grow by 3 per cent to \$78 billion.

The emphasis on basic research echoes the views of leading Republicans in Congress, and contrasts with Clinton's previous emphasis on technology programmes. But Jack Gibbons, the president's science adviser, denies that the administration has been bounced by the Republicans into backing basic research. "We're delighted that there are calls for increased research and development from Capitol Hill," Gibbons says. "It would be fruitless and unproductive to argue about who came first."

The administration proposes to spend the money on research and other investment priorities, including education and the environment, without cutting other spending or consuming the \$9.5 billion budget surplus projected for 1999. The budget avoids breaching the spending limits set in last year's balanced budget agreement by proposing that extra money should be diverted to research through a tobacco settlement.

Colin Macilwain

NIH basks in broad support

President Bill Clinton has asked Congress for the biggest dollar increase in the history of the National Institutes of Health (NIH) biomedical agency, with the focus on cancer research spending.

The request, made in the \$1.73 trillion 1999 budget Clinton sent to the Congress on Monday (2 February), for \$1.15 billion in new money would bring the NIH's budget to \$14.8 billion in 1999 — an 8.4 per cent increase. The budget also projects a 48 per cent increase in spending on NIH by the end of 2003, when the agency's funding would reach \$20.2 billion.

Cancer research fared particularly well, with Vice President Al Gore using a White House ceremony last week to declare scientists "right on the verge" of a breakthrough, and to tout a 65 per cent, \$4.7 billion increase in cancer funding at NIH over the next five years. Ninety per cent of it would go to the National Cancer Institute (NCI).

In spite of this focus on cancer, "everyone is a winner in this budget", says NIH's director, Harold Varmus. "There is not a single activity at the NIH that is not going to be markedly enhanced."

Clinton calls the new money "vital" to America's continued leadership in science and technology, and describes NIH as the "flagship" of his 'Research Fund for America', a \$25.3 billion, five-year package of spending increases on science and technology, of which NIH would ultimately collect \$17 billion.

Like the rest of the fund, the NIH increases

Global warming is bright spot for Energy

Non-military science programmes at the US Department of Energy (DOE) will enjoy a 10 per cent increase under the budget proposal — but much of the money will be consumed by the start of construction at the \$1.3 billion Advanced Neutron Spallation Source at the Oak Ridge National Laboratory in Tennessee (see *Nature* **391**, 421; 1998).

DOE's programme in high-energy physics will get an increase of just 2 per cent to \$690 million, and programmes in biological and environmental research, fusion and nuclear physics will also see little change from their 1998 budgets.

But Martha Krebs, assistant secretary for energy research, says that, within these constraints, the

funds for actually operating DOE's major scientific facilities will expand by \$85 million to \$1 billion.

On the applied research side, the department pledged an extra \$338 million for work to assist in the reduction of carbon emissions. Most of this will be spent on research into renewable energy sources and energy efficiency.

No new funds would be made available for fusion energy, however. After the design phase of the International Thermonuclear Experimental Reactor ends in July, DOE will spend just \$12 million a year on the collaboration, diverting the other \$39 million it was spending on the project to domestic fusion programmes. "We've tried to

maintain our commitment to fusion, but we saw other areas in which we want to invest," says Federico Peña, the energy secretary.

Peña will put an extra \$330 million into the science-based Stockpile Stewardship programme at the department's nuclear weapons laboratories, increasing its budget to \$4.5 billion. He denied that the programme is out of control: "We think the Congress will understand, and we don't anticipate going above the \$4.5 billion mark," he says.

Clinton was due to visit the Los Alamos laboratory with Peña this week to inspect the progress of stockpile stewardship and talk about its relationship to the Comprehensive Test Ban Treaty

C.M.

EPA to spend more on clean air and water

The \$7.8 billion budget request for the Environmental Protection Agency (EPA) is up 6 per cent from this year's level, with much of the increase going to pay for clean water and air programmes. But the agency's overall funding for science and technology goes down 6 per cent, to \$631 million, and the Office of Research and Development drops 9 per cent to \$487 million.

Disaster network boost in increase for USGS

The US Geological Survey's \$807 million budget request represents a 6 per cent increase over last year's appropriation, but funding is projected to decline slightly, to \$796 million, by 2003. The 1999 increase includes money for clean water initiatives, species and habitat conservation studies, and \$15 million for a multi-agency natural disaster information network.

Technology projects fall from presidential grace

This Clinton budget is the first since 1993 to omit grandiose plans for two programmes that were once in the vanguard of the administration's technology policy. No expansion proposals are made for the Advanced Technology Program and the Manufacturing Extension Partnership, both run by the Department of Commerce. □

UK consults public on clones for research

[LONDON] Two advisory groups to the British government are seeking public endorsement of research using cloned human embryos less than 14 days old, pointing out that experiments using conventionally fertilized embryos of this age are already allowed under British law.

The Human Genetics Advisory Commission (HGAC) and the Human Fertilization Embryo Authority (HFEA) also argue that a sharp distinction should be made between 'therapeutic cloning' — a category that would include such experiments — and 'human reproductive cloning', which is already effectively banned in Britain.

A consultation document published in London last week seeks public comment on both aspects of cloning. It is expected to help guide ministers in drafting legislative guidelines setting out the conditions under which experiments with 'therapeutic cloning' would be allowed.

The process is being seen as parallel to that following the birth of the first 'test-tube' baby in 1978. This led to the setting up of the HFEA to police the whole area of infertility research treatment, the hub of a regulatory process described by Sir Colin Campbell, vice-chancellor of the University of Nottingham and chair of the HGAC, as "probably the finest system in the world". Campbell described reproductive cloning as "inefficient, unsafe and morally repugnant". But he warned that it would be unwise to rush into legislation that turned out to be too restrictive both on research and on new forms of infertility treatment. "We are trying to advise ministers and parliament with enormous exactitude on how to frame the regulations," he said.

The decision to produce a document for public consultation was made last year, shortly after the birth of the cloned sheep Dolly. Although the HFEA already has the right to refuse any request to carry out

human cloning — and has indicated that it would do so — it was also felt important to consult on some of the wider issues involved.

In particular, the authority appears keen to head off opposition to research with potential medical applications that would involve the transfer of a nucleus from a mature human cell into an embryonic cell, which might technically be described as human cloning.

Anne McLaren, a member of the HFEA, says that no applications have yet been received for a licence to carry out such experiments, but that this situation might well arise. For example, research which might generate *in vitro* stem cells and cause them to differentiate into specific cell types could provide insights into inducing regeneration of damaged tissue without risk of rejection.

"We expect one day there will be an application for research under 'therapeutic cloning' which will not lead to a new baby, but might lead to medical progress," says Ruth Deech, a family law specialist who is principal of St Anne's College, Oxford, and chair of the HFEA. "We are in a transitional phase, which is why this is a good time to consult the public."

As well as asking for general views about reproductive cloning, the consultation document therefore specifically asks: "Would research using nuclear replacement technology raise any new ethical issues in relation to what is permitted in work with embryos in the 14-day period?"

Groups keen to see the general development of infertility treatments, and which back the government's current endorsement of embryo research under strictly regulated conditions, say they have no difficulties. "I would not see any ethical difference between using an uncloned embryo and a cloned one," says Juliet Tizard of the pressure group Progress.

In contrast, 'pro-life' groups, who remain opposed to all research on fertilized embryos, say they intend to raise the same arguments as were previously used — unsuccessfully — to block more conventional lines of research.

"Ontological arguments do not distinguish between ordinary embryos and embryos that have been cloned," says Peter Garrett of the group Life. "We remain on a collision course [with researchers] on this issue; we continue to argue that the totipotent embryo should not be 'denatured'."

Garrett also argues that giving the green light to therapeutic cloning is likely to lead to an eventual acceptance of reproductive cloning. But Deech disagrees: "I do not think work on therapeutic cloning will make it easier to do reproductive cloning. She adds: "Our experience with embryo research demonstrates that you can take account of public feelings without getting in the way of research."

David Dickson

US sees flurry of bills in bids to legislate

[WASHINGTON] Human embryo research is coming under scrutiny again in the US Congress as lawmakers present various hastily drafted bills to outlaw human cloning.

Richard Arney (Republican, Texas), the House of Representatives Majority Leader, announced last week that the House will move quickly this month to outlaw human cloning in both public and private sectors. Cloning "leads to designer children, organ farms and a growing disregard for the sanctity of life," Arney said. "We intend to stop [cloning] altogether."

The vehicle for Arney's effort, a bill introduced last year by Congressman Vern Ehlers (Republican, Michigan), was being rewritten last week. It is expected to mirror a new Senate bill jointly sponsored by three senators, including Christopher Bond (Republican, Missouri) and Bill Frist (Republican, Tennessee). This bill, which they hoped to introduce on Tuesday (3 February), would include the cloning of human embryos that are not implanted and brought to birth.



Arney: wants to outlaw human cloning completely.

Ehlers, Bond and Frist are promising that their bills will be narrowly targeted to avoid hampering biomedical research. Last year, legislation passed by the House Science Committee explicitly protected the cloning of human cells, tissues, DNA and molecules and the cloning of animals (see *Nature* 388, 505; 1997).

In a separate move, Senators Dianne Feinstein (Democrat, California) and Edward Kennedy (Democrat, Massachusetts) introduced on Monday of this week a more liberal bill. This would essentially codify the recommendations of the National Bioethics Advisory

Commission (NBAC) (see *Nature* 387, 217; 1997 and 387, 644; 1997), except that the bill's moratorium on private- and public-sector cloning for reproductive purposes is doubled in duration to ten years from NBAC's five. This bill explicitly protects the cloning of DNA, molecules, cells, tissue and organs, and the creation of animals by cloning.

The Feinstein-Kennedy bill does not ban the cloning of human embryos that are not implanted and brought to birth, effectively leaving the cloning of embryos for research legal in the private sector. (Each year Congress bans the use of federal funds for human embryo research.)

The emergence of the bills capped a week of intense lobbying in the Congress, with the National Conference of Catholic Bishops and other religious groups pushing for a ban on human embryo cloning, while the Biotechnology Industry Organization (BO) and the American Society for Reproductive Medicine fought to limit any ban to cloning for producing humans. **Meredith Wadman**

Latin America steps up its science tempo

Colin Macilwain

Democracy, price stability and economic liberalization in Latin America are combining to present scientists in the region with their best opportunity in decades to match international standards of excellence.

[BUENOS AIRES] Science in Latin America has reached a crossroads, according to life scientists from Latin America and Canada who met in Buenos Aires, the Argentinian capital, last month. Promising initiatives that are under way could give new impetus to science in the continent — or they could founder, leaving the region dependent on imported science and technology.

But an upbeat mood prevailed at the meeting where the 50 scientists, who receive funding from the US-based Howard Hughes Medical Institute (HHMI), were presenting work carried out in the first year of their five-year Hughes international scholarships.

David Sabatini, an Argentinian-born cell



Sabatini: opportune time to build base.

biologist who advises HHMI on its international programme, says the opportunity for Argentina to develop a strong domestic science base is greater than it has been since he left for the United States in 1961. "The mood is very positive for science," says Glaucius Oliva, who runs a new protein crystallography and structural biology laboratory at the University of São Paulo in Brazil. Oliva says there is "a growing awareness among the general public" that the existence

of a strong science and technology base "is a key separation between the developing and the developed countries".

Scientists in the region, despite national differences, face the same basic level of opportunity. For the first time in 50 years, low inflation is allowing grants to hold their value and real spending on research to increase, while stable, democratic government should allow open and 'clean' peer review to take root. "I don't know if science can ever flourish, except in a democracy," observes Max Cowan, chief scientific officer at HHMI.

Perhaps most auspiciously, domestic industry is being exposed to open competition. As it adapts and updates, it can either invest in research and build better relations with domestic scientific institutes and universities — or bypass them and buy technology from abroad.

This dynamic is leading governments and the international banks that supervise their policies to demand a new level of accountability from science funding agencies. Existing structures — most notably the national councils for science and technology in all the Spanish-speaking countries — are under the microscope, as governments act to ensure the relevance of their work to perceived national needs, and to integrate their research better with teaching at the region's vast, arts-orientated universities.

In Argentina, Jorge Rodriguez, head of the cabinet office and right-hand man to President Carlos Menem, is a US-trained plant geneticist who is implementing far-reaching science policy reforms. The Argentine council for science and technology has been moved from the president's office to the education department, given more money and subjected to an extensive review which, according to government officials, will shape its reform in 1999.

The Chilean president, Eduardo Frei, is himself an engineer, and a key adviser, Claudio Teitelbom, is also orchestrating a review of Chile's council for science and technology. During the past three years 40 new 'president's chairs' have been awarded, supporting top-level researchers to the tune of \$100,000 a year each.

Brazil — the region's largest research and development (R&D) performer by far, with a total annual R&D investment of about US\$4 billion — has just negotiated a loan with the World Bank which will fund \$600 million of applied research, subject to strict, carefully constructed peer review (see *Nature* 391, 317; 1998).

And in Mexico, support for science has grown steadily over the past decade, despite economic difficulties. The government has just introduced an R&D tax credit for industry. Despite growing ties with the United States, Mexican scientists have plenty in common with their South American colleagues and want closer and more formal

'Collaboration could restore excellence'

[BUENOS AIRES] A proposal to set up one or more jointly funded centres for South American life scientists is gathering momentum. It may be discussed at a summit meeting in April of the Chilean, Argentinian and Brazilian presidents.

The centres, which would be along the lines of the European Molecular Biology Laboratory (EMBL) in Heidelberg, are the idea of Argentinian-born David Sabatini, chair of the cell biology department at New York University. Last month in Buenos Aires, Sabatini told a meeting of scientists funded by the Howard Hughes Medical Institute (HHMI) that the centres could help restore South American science to the excellence he experienced when training in Argentina in the 1950s. "I believe we need to look to a supranational institute such as EMBL or CERN [the European laboratory for particle physics]," he says.

But to avoid arguments about location, and build on

existing pockets of strength on the continent, Sabatini proposes more than one centre. He suggests, perhaps, three: structural biology in Brazil, neuroscience in Chile, and molecular biology and biochemistry in Argentina.

Last November, Sabatini took his idea to President Eduardo Frei of Chile, who promised to raise the matter with the leaders of Argentina and Brazil. And in Buenos Aires, Sabatini met Jorge Rodriguez, head of the cabinet office in Argentina, who also expressed support.

Senior officials at HHMI are also enthusiastic. The chief scientific officer, Max Cowan, suggests the institute might consider matching government support for such centres during an initial five-year period. But he stresses that this is a personal view.

Scientists at the meeting were excited by Sabatini's proposal. Enrique Brandan, a cell biologist at the Catholic University of Chile in Santiago, agrees with Sabatini that the centres could develop

existing areas of excellence.

Brandan and several other scientists who were at the meeting also believe the political climate is right for such a venture, with the main South American countries cooperating more closely through Mercosur, their trade group. They think that if the centres gained reputations for quality, they would attract visiting scientists from North America and Europe.

But others question the political will to act jointly. Some, such as Esther Orozco, a geneticist at the National Polytechnic Institute in Mexico City, are keen on the idea but doubt that governments will place a high enough priority on science to pursue the concept.

Other Mexican scientists want to get Mexico involved, even though the country does not belong to Mercosur. Lourival Possani, a biochemist in Mexico City, says he will put the idea to the president's advisory committee on science and technology, of which he is a member. **C.M.**

Argentina gives peer review a boost

[BUENOS AIRES] A new science-funding agency will open in Argentina this year, multiplying by a factor of ten the money available to scientists in peer-reviewed, extramural grants.

Some see the setting up of the agency, however, as a government move to exercise more control over research, and as a threat to their independence.

The agency will aim to establish the peer-review process, according to Juan Carlos del Bello, secretary of science and technology at Argentina's education ministry. "Peer review is not widely applied in Argentina – in fact, it has not really been applied at all," he says.

But one molecular biologist at the national council for science and technology (CONICET), which is currently the main sponsor of basic research in Argentina, argues that peer review within CONICET is already open and fair, and complains that the new agency will be government-controlled, and consequently biased towards technology.

The \$35-million research grant budget of the new National Agency for the Promotion of Science and Technology is modest compared to the \$900 million that the Argentine government will spend on research and development this year. But both the government and the Inter-American Development Bank (IDB), which provided loans to the agency, expect its influence greatly to exceed its raw spending power.



Mariscotti: heading the new agency to promote peer review.

The agency's director is Mario Mariscotti, a respected physicist who is also president of the Argentine National Academy of Sciences. Applications for the agency's first round of grants were invited last August and are now under review. The applications received for the grants – each worth about \$25,000 a year for three years – numbered 2,400, and 700 awards will be made.

The grants are expected to support teams of three people; they exclude salaries, which will normally be paid by the university or institute where the researcher works. The Argentinian government is negotiating with the IDB to increase the value of each grant to an annual \$50,000 next time round.

Fourteen active scientists from different disciplines have been appointed to vet the applications and send them for review by qualified peers inside and outside Argentina. After the best applications are identified, committees of scientists from each of the 14

disciplines will select recipients on the basis of their 'relevance' to national goals as spelled out in the new National Plan for Science and Technology for 1998–2000.

Del Bello, an economist who was responsible for Argentinian universities before taking up his current post in 1996, has harsh words for CONICET. CONICET runs 150 institutes and employs 3,000 staff scientists – leaving only \$4 million of its \$200 million budget for external grants. Although the CONICET institutes are often physically attached to universities, del Bello says the bedfellows have commonly been "in conflict" with each other. The CONICET institutes "are independent entities, isolated and sometimes privileged", he says, and their staff "refuse to teach".

Given such government sentiments about an organization that employs most of the country's leading scientists, it is unsurprising that the new agency has attracted considerable criticism. It is seen by some scientists as an attempt by the government to direct research more forcibly and as a threat to their traditional independence.

The molecular biologist from CONICET says that del Bello wanted to dismantle CONICET when he took office, but was dissuaded from doing so "by scientists' demonstrations". Del Bello denies that he had any such intention – although he points out that the World Bank suggested privatizing the CONICET institutes. **C.M.**

collaboration with them (see box, left).

Most of the scientists who have won Hughes funding completed their PhDs or postdoctoral work in Europe or the United States, and could have prospered there. Yet, according to HHMI officials and the Latin American scientists themselves, the work they presented has some way to go to match that of the Canadian scientists who attended the same meeting.

Many of the scientists see an opportunity

to change that. But they acknowledge that the needs of an élite have to be balanced against the need to build a reasonably broad scientific base. Fernando Reinach, a biochemist at the University of São Paulo and part-time head of the life sciences division of FAPESP, the powerful São Paulo state funding agency, says: "In Brazil, if you just support ten groups, you would kill our scientific base. The question is: how do you build the pyramid?"

Science on the continent faces a moment

of truth in the face of sweeping economic liberalization, says Reinach. The scientific community, he says, can either come out of this turmoil reinvigorated, as a strong player in the economic progress of these countries – or be marginalized as an inefficient machinery incapable of responding to change.

Reinach and Oliva both believe that scientists in Latin America should pay more attention to questions of direct importance to the region. "Scientists need to approach problems that are close to the needs of their own countries," says Oliva. "Some of the good scientists here are too US-orientated," agrees Reinach, who also holds a visiting fellowship at Cornell University in New York. "You don't want to be provincial, but there are unique opportunities here."

He cites the local response to a proposal to sequence the genome of the citrus tree pathogen *Xylella fastidiosa* (see *Nature* 389, 654; 1997). "Craig Venter [the prominent US geneticist] thought it was a great idea," he says. "But a lot of people in Brazil weren't sure it was the right thing to do."

Argentina, with its strong scientific tradition and relatively powerful economy, faces problems typical of the continent. Its huge, under-resourced universities pay scant attention to research, and government officials do not feel that the council for science and technology's 150 research institutes are addressing national needs. Its solution is a new research funding agency (see box).

Scientists face many workaday problems too, the most obvious being the delay of urgently needed supplies, such as enzymes, at customs. "Everything to do with buying supplies is extremely complex," says Raúl Padrón, a structural biologist in Caracas and the sole Venezuelan supported by HHMI. "Things will sometimes be damaged because they are not refrigerated properly at customs. They can take months – or a year."

HHMI hopes to gain more support for science through the leverage of its international scholars programme, which gives 20 Canadians and 27 life scientists from Latin America \$50,000 a year for five years to augment their existing funding.

"This obviously won't solve all the problems," says Purnell Choppin, president of HHMI. "But supporting nine or ten high-quality scientists in a country does two things. It supports these individuals, and keeps them there; and it tells the government of the country that they have top people that an outside organisation such as HHMI is prepared to support."

The governments may now be ready to rise to that challenge. According to Susana Decibe, the minister of education and culture, Argentina wants to double the proportion of gross domestic product spent on R&D, to one per cent. "It's a small figure by the standards of developed countries, but a very ambitious target for us," she says. □

Call for UK science office to return to 'centre of government'

[LONDON] A panel of environmental experts has called on the British government to move the Office of Science and Technology (OST), transferred to the Department of Trade and Industry under the previous Conservative government, back to the Cabinet Office. The call comes from the government's panel of advisers on sustainable development, which says that the importance of technology to society means that the OST needs to be at "the centre of government".

The recommendation appears in the panel's annual report, published today (5 February). Sir Crispin Tickell, chancellor of the University of Kent and the panel's convenor, says that such a move would reinvigorate the Technology Foresight exercise, for which the OST is responsible, and whose results are "uneven".

Euro lab takes the lead in capitalizing on research

[MUNICH] The Heidelberg-based European Molecular Biology Laboratory (EMBL) has become the first international research organization to take a stake in a venture

capital company in exchange for intellectual property based on its research. Following the setting up of the laboratory's Technology Transfer Company, EMBL has accepted 10 per cent of the shares in Lion, a local start-up genomics company, in exchange for rights to its Arakis sequencing system.

Fotis Kafatos, the laboratory's director-general, says that "organizations like ours have the responsibility not only to be innovative, but also to facilitate the use of our discoveries for the benefit of society".

German ombudsman to watch out for fraud

[MUNICH] The University of Konstanz last week became the first university in Germany to appoint an ombudsman for science. The move is in line with the recommendations of a committee of the Deutsche Forschungsgemeinschaft (DFG), the German science council, set up last year to suggest ways of encouraging good scientific practice after a serious case of alleged scientific fraud (see *Nature* 390, 652; 1997).

Rudolf Klein, a 62-year-old professor of physics, has been appointed to the position for two years. According to the university, his role will primarily be one of 'confidant' for young scientists uncertain how to act over suspected scientific misconduct in their laboratories.

Controversial tumour therapy set for trials

[AMSTERDAM] A new European facility that will test a radiotherapy treatment for brain tumours was inaugurated by the research commissioner, Edith Cresson, last week at the European Commission's Joint Research Centre in Petten, the Netherlands. The facility will use the centre's High Flux Reactor. The opening follows a decade of sensitive negotiations with the 15 participating countries.

The treatment, Boron Neutron Capture Therapy, is based on the principle that a neutron beam will selectively destroy tumour cells pre-loaded with boron injected into the bloodstream of patients. But the treatment remains controversial. Previous trials have been relatively unsuccessful, and many experts remain sceptical about the ECU11.4 million (US\$12.3 million) project, funded by the European Commission.

US scientists plea for caution on cloning law

[WASHINGTON] The 52,000-member Federation of American Societies for Experimental Biology has urged the US Congress to ensure that any legislation on cloning should not be permanent, and should be narrowly drawn (see page 523).

But it has not taken a position on whether the cloning of human embryos for research purposes should be outlawed. Last September the federation adopted a voluntary moratorium on human cloning (see *Nature* **389**, 319; 1997).

In a separate move, 55 medical and patient advocacy groups, including the American Academy of Pediatrics, the American Association for Cancer Research and the American College of Medical Genetics, have written to members of Congress asking that any legislation should do "no harm to biomedical research".

British drugs companies in mega-merger plan

[LONDON] The British-based pharmaceutical companies GlaxoWellcome and SmithKline Beecham have announced plans to merge, creating what the two companies describe in a joint statement as "the largest research and development organization in the global healthcare industry".

The new company — expected to be called Glaxo SmithKline — will have a market value of more than £100 billion (US\$150 billion), and a combined worldwide workforce of 110,000. But heavy job losses are expected, particularly among scientists working in areas of research where the two companies' efforts overlap.

Green light for sales of leprosy vaccine in India

[NEW DELHI] The Indian government has authorized the sale of what is claimed to be the world's first leprosy vaccine. It took two decades to develop, and will be made by Cadilla Pharmaceuticals of Ahmadabad. No major side effects were noticed after clinical trials using 80,000 doses.

India has 60 per cent of the world's 1.15 million leprosy cases. Treatment usually entails a three-year course of multiple drugs, with varied results. A course of treatment with the vaccine will last between six months and a year, and will cost US\$1.25. The vaccine will be available from June.

Challenge ahead for boss of Euro research centre

[AMSTERDAM] Herbert Allgeier has been nominated as the new director-general of the European Commission's Joint Research Centre (JRC), which comprises seven institutes in Italy, the Netherlands, Belgium, Germany and Spain. Allgeier, an engineer, currently directs the commission's advanced communication technology programme and coordinates its air and space activities, a position he will retain.

In his new appointment, announced last week, Allgeier will have to respond to

increasing pressure, particularly from the European Parliament, to reorganize the way in which the JRC's institutes are run and to increase the quantity and quality of their scientific output.

Setting sail to study underwater volcano

[SEATTLE] Oregon State University is to send a research ship to investigate an underwater volcano, Axial Seamount, that has been erupting off the Oregon coast for the past week. Chris Fox, a geologist with the National Oceanic and Atmospheric Administration, says the eruption has not been seen. But he says it is likely that rivers of red-hot lava are flowing out of the volcano, along with giant plumes of scalding, mineral-rich water carrying microbes that thrive beneath the ocean floor. The volcano is 4,500 feet high. Its peak is nearly 4,000 feet below the ocean surface.

Correction

In last week's article on New Zealand (*Nature* **391**, 426; 1998), an editing error led to the statement that, in the five years to 1995–96, the number of support staff in Crown Research institutes "tripled to more than 900". This should have read: "increased by about 350 to more than 900".

Place of science on European agenda

Sir — Your leading article of 1 January (*Nature* 391, 1; 1998) voices warranted concern over the fate of fundamental research endeavours of the European Union, which you trace back to the Rome Treaty of 1957. But you erroneously credit the treaty with the creation of the European Commission, which was created only in 1968.

The 1957 treaty created two institutions. One, considered the most important at the time because of the Suez oil crisis, was the European Atomic Energy Community (or Euratom), with its Euratom Commission. The other was the European Economic Community (also known as the Common Market), with its Economic Commission. In 1968, the two communities merged, together with the pre-existing Coal and Steel Community, to form the European Community, headed by the single European Commission.

So, from 1957 to 1968, there was a separate Euratom Community, with its own commission, to which the Rome Treaty had

assigned the creation of a European nuclear industry. Despite the industrial goal, or rather because of its obvious requirements, fundamental science was not absent from Euratom directives. Research and education are written into the treaty, which even includes the creation of a European University. In fact, a large part of the research enterprise at that time was connected to nuclear establishments, in the United States, in other developed countries and in developing countries.

Euratom research and development was carried out in two ways. On the one hand, a Euratom Common Research Centre was set up, including the research establishment at Ispra, Italy, where the Orgel prototype reactor was designed and built, the Bureau of Nuclear Standards at Mol, Belgium, the European Trans-Uranium Institute in Karlsruhe, Germany, and the research reactor at Petten, The Netherlands.

On the other hand, Euratom created associations with national institutions (public or private) in the realm of nuclear

reactor prototypes, fusion (culminating in the Joint European Torus, JET, project), isotopic geology, radiobiology and so on. There was also a major US-Euratom agreement with an important research and development component. The teams were multinational, and the knowhow and results were made available to all member states. The Common Research Centre is still active, as are some major associations (JET, for example).

So, when the European Commission was created, in 1968, it is not so much that, as you say, "science was unfortunately not on their agenda". Rather, science was dropped from the primary agenda, after 11 years of turmoil and opposition to an inspired and in many ways successful European research and development programme (see, for example, *Sciences et Avenir*, pp 214–217, March 1969).

Maurice Guéron

BIOP, Ecole Polytechnique,
Palaiseau, France
e-mail: mg@pmc.polytechnique.fr

Order and justice in French universities

Sir — Your report "Battle to boost the status of French universities" rightly points out that French universities have been the poor relation of the *grandes écoles* (*Nature* 391, 6; 1998). This is correct in terms of economic evaluation and may be explained by the leading administrative positions occupied by former students of these *écoles*. I take exception, however, to your statement and tone when you say that "until now French universities have played second fiddle in science to the public research organizations" and that "university staff often have almost no time for research".

To begin with, all French researchers are educated and trained in the universities before they enter research laboratories located, mostly, on university campuses, less frequently in public research organization buildings and almost never in the *grandes écoles*. Second, these trainees have undergone ruthless selection processes in the university graduate schools, where fewer than 5% of initial graduates are accepted for PhD programmes. Third, at every stage of their training, these future researchers are 'supervised' by university staff who teach through their own example the way in which research and education in science are intimately associated.

Finally, it should be recognized that the great majority of laboratories at any of the

leading public research institutes in France are directed not only by university graduates but more strikingly by university staff and professors. That these talented teachers have been operating in less than favourable conditions when compared to their more fortunate colleagues from the other research organizations is certainly true and this is where it is to be hoped that the education minister, Claude Allègre, will introduce order and justice.

A. D. Strosberg

Institut Cochin de Génétique Moléculaire,
UPR-0415 du CNRS,
22 rue Méchain, 75014 Paris, France
e-mail: strosberg@icgm.cochin.inserm.fr

Risk and benefit

Sir — The assessment of any new technology demands a balanced consideration of both the possible associated risks and the perceived social benefits. All too often, however, regulatory apparatus provides only for the scientific assessment of risk, thus concealing the corresponding consideration of wider issues.

We note that, under the provisions of the Convention on Biological Diversity, negotiations about an international 'Biosafety Protocol' for the regulation of the use and transfer of genetically modified organisms are approaching conclusion. Whether or not this protocol will provide for the consideration of social and

economic issues is as yet undecided.

As individuals involved in either the development or implementation of risk-assessment frameworks for biotechnology, we are unhappy with the unspoken expectation that we should assess not just the possible risks associated with the commercial release of a genetically modified organism but also, tacitly, the broader social implications. We submit, therefore, that new regulations should permit the overt consideration of socio-economic issues before the environmental release or transboundary movement of genetically modified organisms.

Implementation of such provisions may not be straightforward, but their inclusion would at least allow open discussion of issues that many regulators are currently obliged to consider implicitly.

Tom Crompton

UNIDO/ICGEB Liaison Office,
Vienna International Centre, Austria
e-mail: tcrompton@unido.org

John Barrett

Department of Genetics,
University of Cambridge, UK

Chris Gliddon

University of Wales, Bangor, UK

Julian Kinderlerer

Sheffield Institute of Biotechnological
Law and Ethics, University of Sheffield, UK

George Tzotzos

UNIDO Investment and Technology
Promotion Division,
Technology Service, Vienna, Austria

Animal embryos in deep time

Stefan Bengtson

Discoveries of spectacularly preserved embryos and tissues, in rocks that are about 570 million years old, open a new era in the study of early animal evolution.

A spell seems to have been broken — animals considerably older than the Cambrian are finally being found in the fossil record, and they are preserved in a way that reveals details down to the cellular level. This stirring claim is based on studies of the roughly 570-million-year-old Doushantuo phosphorites in southern China, and is hardly weakened by the fact that it comes from two different groups of palaeontologists. On page 553 of this issue¹, Xiao *et al.* report on exquisitely preserved algae and animal embryos from the phosphorites, while, in *Science*, Li *et al.*² describe sponges and animal embryos from the same deposits.

The Cambrian Explosion, the evolutionary radiation of life forms about 550 million years ago, was one of the major turning points in the history of the Earth. Over a period of a few tens of millions of years, practically all of the principal animal lineages (phyla) appeared. At least this is what the fossil record suggests. The dearth of animal fossils below the Precambrian–Cambrian boundary has been one of the main frustrations in studies of early animal history, for there has been very little agreement on how, when and where animals lived and evolved during the Proterozoic — the period of time that ended with the beginning of the Cambrian and the Phanerozoic (see Fig. 1 for a timescale). The celebrated Ediacara biota of the Vendian Period, mostly preserved as impressions and traces in sandstones and shales, provides some clues. But it has so far produced more disputes than data with regard to early animals; most Ediacaran deposits are in any case only slightly older than the Cambrian³.

For a long time following Darwin, a long interval of hidden animal evolution before the Cambrian used to be assumed. But owing to the influence of Preston Cloud^{4,5} and others, the more recent inclination has been not to postulate more hidden evolution than absolutely necessary. An estimate in 1996 by three palaeontologists suggested that the main divergence of animal phyla occurred no earlier than 565 million years ago⁶. At the same time, an attempt⁷ to estimate the same divergence times using sequence comparisons of various animal genes landed at figures that were more than twice as high. This

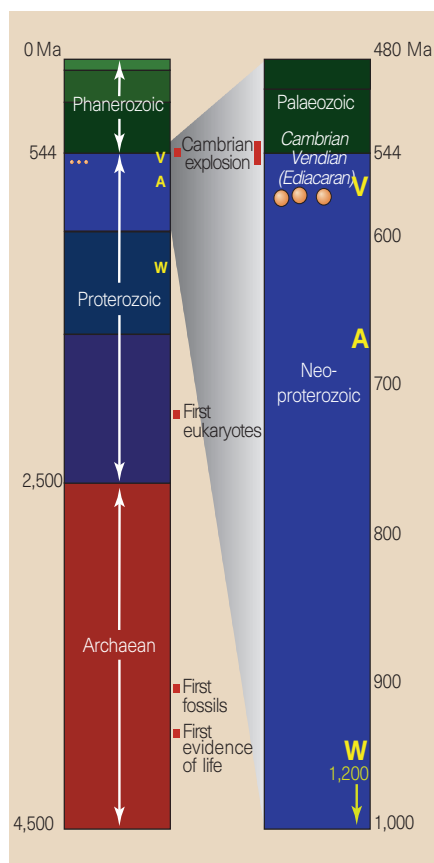


Figure 1 Timescale of Earth's history (left), and the interval between 1,000 and 480 million years ago (Ma), showing some significant events in the fossil record. The Doushantuo fossils described by Xiao *et al.*¹ and Li *et al.*² are indicated by three spherules. V, A and W show the alternative approximate ages (without error bars) of the main radiation of animals as proposed by Valentine *et al.*⁶, Ayala *et al.*⁸ and Wray *et al.*⁷. Specifically, these are alternative dates for the divergence between protostomes and deuterostomes.

analysis has been criticized on methodological grounds, and a date of about 670 million years⁸ has been proposed instead. Although most palaeontologists will probably be quite happy with the latter date, the fossil record has hitherto been ominously silent.

The origin of animals has thus been as elusive as that of life itself: there are many possibilities but few facts to test them against. The Doushantuo fossils promise to

change all that — not only do they give the first convincing glimpse of pre-Ediacaran animal life, but the quality of preservation is almost unheard of, even in much younger fossils. This preservation offers insights into cell-level anatomy, embryological development and life cycles — such matters have not normally been considered to be open to investigation in fossils. Above all, the usual explanation for the missing fossil record has been that the animals were soft and small and therefore would not be preserved as fossils. Now such fossils are beginning to appear. Even better, there seems to be nothing very unusual about the Doushantuo phosphorites, and they may therefore show the way to many other such sites.

The key to the exquisite preservation is calcium phosphate. This mineral is known for its faithful replication of delicate tissues^{9,10}, although early phosphatization in sediments tends to be very patchy and frequently only fossils of millimetre size or below are preserved^{11,12}.

The specimens described by Li *et al.*² are interpreted as sponges. The siliceous or calcareous spicules of sponges are common in the fossil record, and sponges are among the simplest of multicellular creatures. So they would not be unexpected members of the earliest animal assemblages. The needle-shaped spicules in the examples depicted by Li *et al.* are regularly arranged in distinct bodies built up of cell-like objects, some of which adhere to the spicules in the same way as sclerocytes (spicule-forming cells) do in living sponges. Details of the proposed interpretations are open to question, but it will be difficult to disprove that these are indeed sponges.

The Doushantuo sponges are small, about 150–750 μm in maximum dimension. Although Li *et al.* interpret them as fully formed adults, they could also represent propagules or embryos. But the authors quite reasonably interpret other fossils as embryos of different kinds of animals, and here they find themselves in good agreement with Xiao *et al.*¹, who have isolated a suite of globular fossils from the Doushantuo which they identify as animal embryos in the early stages of cell division (cleavage).

Animal embryos are small and delicate. A few years ago, the thought of finding fossilized embryos of anything but bony pre-hatchlings of dinosaurs and the like was preposterous. However, recent discoveries of Cambrian phosphatized embryos of animals in various developmental stages^{13,14} suggested that this might be a fruitful search strategy for the missing record of Proterozoic animals¹⁴. It indeed seems that all we needed was to open our eyes to the possibility: the fossils now identified as embryos had actually been described in the literature but were interpreted as colonial green algae¹⁵.

Xiao and colleagues' fossils are about half

a millimetre in diameter, and are compartmentalized into two, four, eight or more bodies which are proposed to be blastomeres (cells in a cleavage embryo; see the stunning picture on the cover and those on page 556 of this issue). The constant size of the fossils, irrespective of the number of compartments, fits a pattern of developing early embryos with a constant cytoplasmic volume. This would not be expected in colonial algae or in objects formed by non-biological processes.

So what information can we get from these kinds of fossils? The two studies^{1,2}, although preliminary in nature, already yield some insights. The previously known oldest sponges were late Ediacaran hexactinellids (glass sponges)^{16,17}, but the spicule morphology and cell configuration of the Doushantuo sponges are very different from those of hexactinellids. The early-cleavage embryos have a tetrahedral blastomere configuration that is today known in some animals such as nematodes, flatworms and arthropods. One should be careful about drawing evolutionary conclusions from physically simple patterns like these, but the observation clearly shows the potential of such material. Only finds of later developmental stages will tell in which direction, and how far, these embryos developed.

The Doushantuo phosphorites cover an area of 57 km², and they undoubtedly contain further secrets. Phosphate is thus pay

dirt. But whereas digging up the basal roots of animals may have its particular appeal, let's not forget about the rest of animal history. Developmental and evolutionary biology are complementary but largely separate sciences, and the fossil record might help in bringing them together. Palaeoembryology may be a science of the past, but it could have a brilliant future. □

Stefan Bengtson is in the Department of Palaeozoology, Swedish Museum of Natural History, Box 50007, S-104 05 Stockholm, Sweden.
e-mail: Stefan.Bengtson@nrm.se

1. Xiao, S., Zhang, Y. & Knoll, A. H. *Nature* **391**, 553–558 (1998).
2. Li, C.-w., Chen, J.-y. & Hua, T.-e. *Science* **279**, 879–882 (1998).
3. Grotzinger, J. P., Bowring, S. A., Saylor, B. Z. & Kaufman, A. J. *Science* **270**, 598–604 (1995).
4. Cloud, P. E. *Evolution* **2**, 322–350 (1948).
5. Cloud, P. E. Jr in *Evolution and Environment* (ed. Drake, E. T.) 1–72 (Yale Univ. Press, New Haven, CT, 1968).
6. Valentine, J. W., Erwin, D. H. & Jablonski, D. *Dev. Biol.* **173**, 373–381 (1996).
7. Wray, G. A., Levinson, J. S. & Shapiro, L. H. *Science* **274**, 568–573 (1996).
8. Ayala, F. J., Rzhetsky, A. & Ayala, F. J. *Proc. Natl Acad. Sci. USA* **95**, 606–611 (1998).
9. Müller, K. J. *Phil. Trans. R. Soc. Lond. B* **311**, 67–73 (1985).
10. Martill, D. M. *Nature* **346**, 171–172 (1990).
11. Walossek, D. *Fossils & Strata* **32**, 1–202 (1993).
12. Briggs, D. E. G. & Wilby, P. R. *J. Geol. Soc. Lond.* **153**, 665–668 (1996).
13. Zhang, X.-g. & Pratt, B. *Science* **266**, 637–639 (1994).
14. Bengtson, S. & Yue, Z. *Science* **277**, 1645–1648 (1997).
15. Xue Y.-s., Tang T.-f., Yu C.-l. & Zhou C.-m. *Acta Palaeontol. Sinica* **34**, 688–706 (1995).
16. Gehling, J. & Rigby, J. K. *J. Paleontol.* **70**, 185–195 (1996).
17. Brasier, M., Green, O. & Shields, G. *Geology* **25**, 303–306 (1997).

Tracer studies using highly surface-active metal isotopes show that the marine colloidal phase is very dynamic, with high colloid aggregation rates being sustained even in nutrient-poor surface waters^{8,9}. These findings imply that colloid production rates must also be high, presumably through a combination of cell exudation and lysis, microbial degradation of particulate organic matter, and 'sloppy' feeding and excretion by zooplankton^{1,10}. Colloidal concentrates obtained by cross-flow ultrafiltration contain carbohydrates, proteins and lipids³, as well as significant quantities of the biologically essential metals iron, copper, zinc, nickel and cadmium⁵.

So what factors control colloid residence times in sea water? Most recently, interest has largely centred on the biodegradation removal pathway. Indeed, size-fractionation experiments suggest that colloidal organic matter supports the bulk of heterotrophic microbial production in sea water, rather than truly soluble substances¹¹. In this case, colloid 'removal' rates would be primarily a function of enzyme-specific reactions which, in turn, would be strongly influenced by the composition and steric accessibility of molecules within colloidal matrices.

However, the results of Chin *et al.*⁶ show the importance of nonspecific surface interactions in colloid cycling. They find that discrete natural polymers coalesce spontaneously to form large, sinking conglomerates (Fig. 1); from the results, it seems that the process has second-order kinetics, with aggregation reaching apparent equilibrium after about 80 hours under the authors' abiotic test conditions. This period is similar to estimates of colloid turnover times in surface waters from ²³⁴Th modelling^{8,9}. Ion-bridging between biopolymers seems to be the primary driving mechanism because the process is reversed by chelation of Ca²⁺ and Mg²⁺. As with colloidal concentrates, histological staining indicates that the aggregates are composed of carbohydrates, proteins and lipids — which is particularly interesting, because it suggests that the stability of a wide range of biocompounds within the dissolved-organic-matter component of sea water may depend less on their chemical composition than on their interfacial characteristics.

Chin *et al.* go on to show that these aggregates undergo reversible swelling and condensation as a function of hydration, a behaviour characteristic of gels formed by tangled networks of polymers. Although similar behaviour has been shown for terrestrial humic matter¹², this is the first report for marine aggregates. These changes were observed under chemical conditions very different from those characteristic of sea water, but Chin *et al.* suggest that similar transitions might occur over natural ranges of temperature and pressure or in chemically

Marine colloids

A neglected dimension

Mark L. Wells

Marine colloids are the most abundant particles in the oceans^{1,2}. They account for 30–50% of the 'dissolved' organic carbon in sea water^{3,4}, and they also contain biologically essential metals⁵. Although great strides have been made in assessing the bulk composition of the marine colloidal phase, there is little consensus about what processes regulate colloid cycling, or even what the ultimate fate of these substances is. The work of Chin *et al.* (page 568 of this issue⁶) adds a fresh perspective to these basic questions by invoking polymer gel theory to explain the behaviour of marine colloids. This insight is important because emerging evidence suggests that the marine colloidal phase is intrinsic to a variety of oceanographic processes, with potential ramifications running from the modulation of metal ion uptake by algae to global climate change.

Colloids lie at the boundary between soluble chemical species and sinking particles, and are defined as substances sized between about 1 and 1,000 nm in diameter. Particles in this range are large enough to acquire an

interface (that is, the interior is chemically distinct from the surrounding medium) but small enough for gravity not to be the dominant force acting upon them. So removal of colloidal substances from sea water depends upon either their degradation to soluble components or their aggregation to form sinking macroparticles.

It has long been recognized that colloids strongly affect the behaviour of carbon and metals in estuaries⁷. But their role in coastal and offshore sea waters had been largely overlooked, mainly because oceanographers have traditionally employed filters of 0.2–0.7 µm pore size to arbitrarily partition sea water into 'dissolved' and 'particulate' phases. Only recently have concerted efforts begun to discriminate colloidal from truly soluble components within the 'dissolved' phase. Given the early evidence that the marine colloidal phase may account for upwards of 250 gigatonnes of potentially bioreactive organic carbon, it is not surprising that new emphasis is being placed on understanding the sources, cycling and fate of these substances.

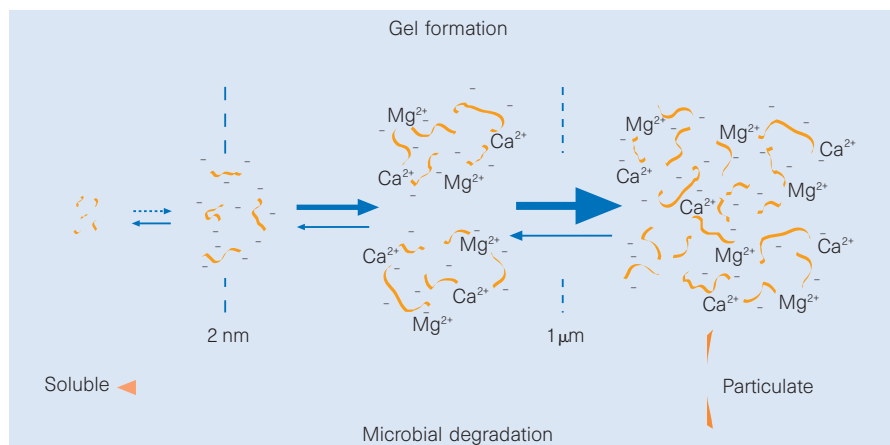


Figure 1 Representations of the dynamic interplay between marine gel formation and microbial degradation in sea water. The negative charges depict the anionic nature of marine organic matter, while cation bridging appears to drive polymer coalescence. Thickness of arrows indicates relative reaction rates. Previous tracer experiments in ocean surface waters had implied that colloidal-sized organic phases could aggregate quickly, but Chin *et al.*⁶ are the first to directly quantify colloid aggregation rates in sea water. They invoke polymer gel theory to explain the mechanism of aggregate formation as well as the physicochemical nature of the resultant macroparticulate phases.

distinct micro-environments. If so, gel condensation might encase biologically labile molecules within a tangled network cage that is sterically impervious to enzymatic attack. Given the swiftness of gel condensation, marine macrogels might be prominent precursors for the geopolymer condensates that are thought to help preserve organic matter in some marine sediments¹³.

Assuming basic polymer theory is well suited to the heterogeneous cornucopia of organic constituents in sea water, it could provide a pleasingly straightforward explanation for the old apparent ¹⁴C ages of oceanic dissolved organic carbon (2,500–5,900 years before present)¹⁴. Because the probability of polymer gels coalescing and becoming stable increases in proportion to the square of polymer length, the smallest organic constituents in sea water would be less likely to be incorporated into macropolymer assemblies than would colloidal constituents. Indeed, recent data indicate that colloidal organic carbon in the upper water column is only a few decades old, whereas the soluble dissolved organic carbon dates to 4,000 years or more¹⁵.

Perhaps the most startling aspect of Chin and colleagues' results is that particulate gels re-form from soluble and colloidal precursors after previous gels are removed by gentle filtration. The authors argue that the initial macrogels do not disintegrate upon filtration but that new macrogels form as a consequence of re-equilibration in the system. If confirmed, this finding has serious implications for oceanographers because it means 'dissolved' (that is, filtered) samples will, in time, come to contain particulates again. More strikingly, it raises the possibility that colloid aggregation in sea water may be strongly influenced by particulate removal rates.

The main component missing in these

experiments is biology, so we cannot assess the relative importance of competing processes of biodegradation and macrogel formation (Fig. 1). The balance of this tug-of-war will almost certainly vary for different substances and over time, because biodegradation may diminish the chance that residues will enter the aggregation cascade. Moreover, phytoplankton and microbial production are not coupled rigidly, so

input of colloids may outpace degradation over short periods. One example may be the apparently spontaneous appearance of macropolymers responsible for the rapid aggregation of marine snow¹⁶. Although the debate over the dynamics and fate of colloidal matter in sea water continues, Chin *et al.* contribute a rich new framework for examining these basic questions. □

Mark L. Wells is at the Institute of Marine Sciences, University of California Santa Cruz, Santa Cruz, California 95064, USA.

e-mail: mwells@cats.ucsc.edu

1. Koike, I. S., Hara, S., Terauchi, K. & Kogee, K. *Nature* **345**, 242–244 (1990).
2. Wells, M. L. & Goldberg, E. D. *Nature* **353**, 342–344 (1991).
3. Benner, R., Pakulski, J. D., McCarthy, M., Hedges, J. I. & Hatcher, P. G. *Science* **255**, 1561–1564 (1992).
4. Guo, L., Coleman, C. H. & Santschi, P. H. *Mar. Chem.* **45**, 105–119 (1994).
5. Moran, S. B., Yeats, P. A. & Balls, P. W. *Cont. Shelf Res.* **16**, 397–408 (1996).
6. Chin, W.-C., Orellana, M. V. & Verdugo, P. *Nature* **391**, 568–572 (1998).
7. Sholkovitz, E. R., Boyle, E. A. & Price, N. B. *Earth Planet. Sci. Lett.* **40**, 130–136 (1978).
8. Moran, S. B. & Buesseler, K. O. *Nature* **359**, 221–223 (1992).
9. Baskaran, M., Santschi, P. H., Benoit, G. & Honeyman, B. D. *Geochim. Cosmochim. Acta* **56**, 3375–3388 (1992).
10. Wells, M. L. & Goldberg, E. D. *Limnol. Oceanogr.* **39**, 286–302 (1994).
11. Amon, R. M. W. & Benner, R. *Limnol. Oceanogr.* **41**, 41–51 (1996).
12. Benedetti, M. F., Van Riemsdijk, W. H. & Koopal, L. K. *Environ. Sci. Technol.* **30**, 1805–1813 (1996).
13. Hedges, J. I. & Keil, R. G. *Mar. Chem.* **49**, 81–115 (1995).
14. Bauer, J. E., Williams, P. M. & Druffel, E. R. M. *Nature* **357**, 667–670 (1992).
15. Santschi, P. H. *et al.* *Geochim. Cosmochim. Acta* **59**, 625–631 (1995).
16. Passow, U. & Aldredge, A. *Deep-Sea Res.* **42**, 99–109 (1995).

Immunodeficiency viruses

1959 and all that

Simon Wain-Hobson

Origins continue to fascinate. When it comes to biology, Darwin set the stage and interpreted the play. Following the revolution in molecular biology, we know how to decode the digital information stored in DNA, and can look back in wonder at the phylogenetic trees of protein sequences such as cytochrome *c* or α -globin. Phylogenetic analysis was a fairly scholarly occupation until the polymerase chain reaction (PCR) discombobulated us all in the late 1980s. PCR allows billion-fold (and more) amplification of a precise segment of DNA — single DNA molecules can be analysed, and almost every living thing has had some part of its DNA amplified by now. Even Neanderthal bones in a museum drawer have harboured enough DNA to yield to PCR, so bringing new life to an old debate¹. And laboratory freezers the world over contain old serum samples, from Africa and elsewhere. Which brings us back once again to the human immunodeficiency virus, and to the paper by Zhu *et al.*² on page 594 of this issue.

Where did HIV come from? Both of the AIDS viruses, HIV-1 and HIV-2, originated

in Africa, where the spread of AIDS preceded development of the disease elsewhere on the globe. How? As is often the case with microbes, a jump from one species to another is probably to blame. This conclusion comes from the observation of viruses with qualitatively identical genetic make-ups among chimpanzees (for HIV-1) and sooty mangabeys (for HIV-2), in geographically overlapping regions. Of the two viruses, HIV-1 is wreaking global havoc — although HIV-2 is capable of generating an AIDS epidemic, HIV-1 has a faster disease course and more efficient transmission.

When did the AIDS epidemic begin? There is the feeling that HIV first lingered in some people before going global. If so, then how long has it been around in humans? That's a more tricky question. To get some idea, researchers have turned to sera and tissue samples collected many years ago for other studies. Artefacts can crop up when using old sera in immunological tests. However, sera may also harbour viruses such as HIV. This is precisely what Zhu *et al.*² have found in a sample drawn back in 1959 in

Leopoldville, which was then part of the Belgian Congo and is now Kinshasa in the Democratic Republic of Congo.

The HIV virion harbours an RNA genome. Molecular biologists prefer to manipulate DNA — it's easier. Hence, to get at HIV's genetic information, Zhu *et al.* extracted viral RNA from plasma and then copied it into DNA, in an *in vitro* reaction using reverse transcriptase. With PCR it is usually child's play to amplify a few thousand nucleotides. As it turned out, the HIV genomic RNA was substantially degraded, meaning that, after much labour, the authors could obtain only a few 300-nucleotide fragments of DNA. Fortunately, the genetic variation of HIV is so great that even with short segments there is ample information to allow comparisons with the enormous HIV database³, and construction of robust phylogenetic trees.

Where does this 1959 virus sequence, designated ZR59, sit with respect to other African isolates? Comparing contemporary HIV sequences reveals a little Big Bang, a radiation in HIV sequence space (Fig. 1).

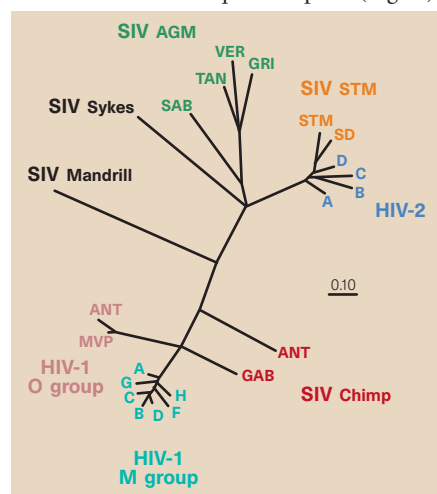


Figure 1 Phylogenetic tree of extant human and simian immunodeficiency viruses. The distance between any two viral isolates is found by following the shortest path relating them (units are arbitrary on a scale from 0 to 1). The global AIDS epidemic is due to the HIV-1 M group of viruses. The O group of HIV-1 viruses also cause AIDS, but they are found mainly in west central Africa. The isogenic chimpanzee viruses (SIV Chimp) are apathogenic (so far). HIV-2 and the isogenic sooty mangabey group (SIV STM) cluster closely. The African green monkey virus group (SIV AGM) is quite diverse. Other monkey viruses are those from mandrill (SIV Mandrill) and Sykes monkey (SIV Sykes). The zero-length branch points (HIV-1 M, O and SIV Chimp, as well as HIV-2, SIV AGM and SIV Sykes) must be considered uncertain for the time being. The tree was constructed from the complete gag nucleotide sequences, or 676 bases after elimination of insertions and deletions. For methods see ref. 4. (Figure produced by Thomas Leitner; e-mail: tkl@t10.lanl.gov).

The sequences map near the edge of this radiation, the distance from the centre being roughly proportional to time from the beginning of the AIDS epidemic. Not so ZR59, which is close to the centre, suggesting that it is but a few nucleotides and years from the virus that founded the epidemic — perhaps no more than 10–15 years, given what we know about the rate of accumulation of substitutions in the HIV genome. This all makes good sense.

What else is the position of ZR59 among HIVs telling us? First, it probably means that the global epidemic was indeed founded by a single HIV although, in this respect, it is no different from the annual 'flu strain. Second, the centre of the radiation and ZR59 are a considerable stretch from any simian counterpart, suggesting that HIV had a human history before it went global. Third, the Big Bang seems to have occurred around, or just after, the Second World War. Emerging

microbial infections often result from adaptation to changing ecological niches and habits. And, of course, the post-war era saw the collapse of European colonialism and attendant changes in urban and technological traits. As usual, when data are limited we're in the realm of speculation, meaning that the story is not over. Nevertheless, this is another powerful demonstration of molecular sleuthing thanks to PCR.

In 1959, the Nobel prize for physiology or medicine was awarded to Severo Ochoa and Arthur Kornberg for their work on nucleic-acid polymerases, while the world rocked around to Elvis and Chuck Berry. There was fog in the English Channel. □

Simon Wain-Hobson is at the Institut Pasteur, 28 rue du Dr Roux, 75724 Paris Cedex 15, France.
e-mail: simon@pasteur.fr

1. Krings, M. *et al.* *Cell* **90**, 19–30 (1997).
2. Zhu, T. *et al.* *Nature* **391**, 594–597 (1998).
3. <http://hiv-web.lanl.gov>.
4. Leitner, T. *et al.* *J. Virol.* **71**, 4761–4770 (1997).

Nanocrystals

The strongest size

Sidney Yip

How crystal grains deform in a material under stress is important both scientifically, for the understanding of plastic flow in solids, and technologically, for its direct effect on the material strength. Grain size plays an important part: in polycrystals with grain sizes in the micrometre range, strength increases with decreasing size. This is known as the Hall–Petch effect^{1,2}, and was first explained in terms of the piling up of dislocations, created by the shearing of crystal planes in each grain, at the grain boundaries. As the grains become smaller, the effect of dislocation blocking increases, thereby strengthening the material. But with the synthesis of nanocrystals, with grains in the nanometre range, the opposite behaviour was found — specimens of copper and palladium softened with decreasing size³.

Why? As reported on page 561 of this issue⁴, computer simulations show that the reverse Hall–Petch effect arises primarily from sliding motions at grain boundaries. This is a timely demonstration of how materials simulations can now provide atomic-level insights into a wide range of physical phenomena.

Schiøtz *et al.*⁴ simulated the molecular dynamics of crystals about 10 nm on a side, with grain sizes between 3.3 and 6.6 nm, using an interatomic potential model appropriate for copper. The simulations amount to a series of virtual tensile-strength tests at very low temperature, stretching the crystal sample to see how individual atoms move. The position of every atom in the sample during the course of deformation is determined and recorded, along with the forces

they exert on each other, so the average stress (tension) in the sample can be found at every step of strain (deformation). The stress–strain variation shows the expected yield and plastic flow, and when the results are correlated with grain size, they indicate a reverse Hall–Petch effect — a stress reduction with decreasing size. By analysing local atomic environments and visualizing individual atomic motions, the authors demonstrate that the deformation stems mainly from many small independent slip events in the grain boundaries.

It is not surprising that the variation of strength or hardness with grain size should have a characteristic length scale. When a polycrystal undergoes large-strain deformation, intragranular and intergranular mechanisms are activated and in general will compete with each other. As each has a different dependency on grain size, there will be a transition region where one takes over from the other. The behaviour is 'normal' when grain size d is greater than the characteristic length⁵, d_c (Fig. 1). The reverse effect, when $d < d_c$, signifies a change in the dominant mechanism of deformation.

No theory yet describes the entire range of behaviour depicted in Fig. 1, but an estimate of d_c can be made if one assumes that it is roughly the smallest grain size that can sustain a dislocation pile-up, and take this to be the equilibrium distance between two edge dislocations⁵. Below this size, dislocation blocking should break down, and intergranular sliding should become relatively important (as the proportion of surface atoms increases). Using these assumptions, d_c for copper and

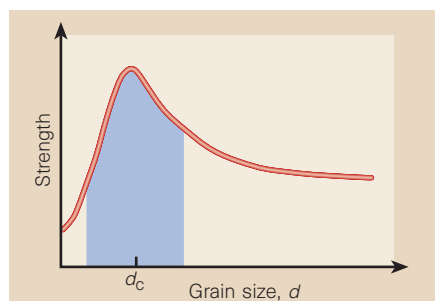


Figure 1 Variation of strength with grain size for metals. For grain sizes of a few nanometres (region shaded blue), strength first increases and then decreases with decreasing grain size. The latter effect results mostly from small-scale sliding events in the grain boundaries⁴. (Adapted from ref. 5.)

palladium is 19.3 nm and 11.2 nm, respectively. These values are quite consistent with the behaviour found by Schiøtz *et al.*⁴, as their simulation grain sizes are smaller.

The practical significance of this behaviour is in the design of new materials through manipulation of microstructure — a long-standing goal of materials science and technology. A recent success in this field is the synthesis of a tough ceramic reported by Chen and Rosenflanz⁶. In this case, a solid solution of α -Si₃N₄ with a whisker-like microstructure was produced which is 40% harder than the β -phase material currently in use, and just as strong and tough. Similarly, hardness could be optimized by controlling grain size.

A factor that is not shown in Fig. 1 is the effect of temperature on material strength. For instance, diffusional creep becomes important at high temperatures. Creep is a difficult phenomenon to study by simulation because the strain rate involved is so slow. But here also, a size effect comes into play: whereas diffusion through the lattice (strain rate proportional to d^{-2}) is indeed not currently accessible, the process of diffusion in the grain boundaries (Coble creep, strain rate proportional to d^{-3}) could be studied by simulation under favourable conditions⁷. For nanocrystalline silicon of grain size 5 nm at two-thirds melting temperature, the strain rate for Coble creep is estimated to be about 10^{-7} s^{-1} , which is within the range of current simulation capabilities. This is another example of the emerging capabilities of simulation.

The behaviour of granular materials near d_c is particularly interesting because of the crossover, a continuous transition from one type of mechanism to another. We have seen that, in the nanocrystal context, as grain size is reduced across d_c , the deformation mechanism changes from dislocation nucleation and motion to small-scale sliding in the grain boundaries — in other words, from an intra-granular process to an intergranular one, or from bulk to interface. A similar transition

may occur in nanometre-scale electronic components, such as epitaxial films and quantum dots. In these heterostructures, the lattice mismatch between different materials induces strain, and how it relaxes during fabrication will determine the material structure, and hence its properties. Again, there is competition between processes mediated by misfit defects such as dislocations, and surface-based processes such as roughening or morphological change⁸.

Simulations can provide details of deformation processes at the atomic level that are not accessible by any other means. Advances in computing have made it possible to study systems of hundreds of millions of atoms, enough to represent physical structures on the nanometre scale or larger. Recent studies have coupled these atomic-scale simulations to finite-element techniques to investigate dislocations⁹. So we can already model material properties across several length scales, from electronic structure to atoms to

microstructure to macroscopic scales¹⁰. In the future, we will also be able to link different simulation methods in order to treat long-range elastic interactions and the local effects of chemical bonding on the same footing. □

Sidney Yip is in the Nuclear Engineering Department, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.
e-mail: syip@mit.edu

1. Hall, E. O. *Proc. Phys. Soc. Lond. B* **64**, 747–753 (1951).
2. Petch, N. J. *J. Iron Steel Inst.* **174**, 25–28 (1953).
3. Chokshi, A. H., Rosen, A., Karch, J. & Gleiter, H. *Scripta Metall.* **23**, 1679–1684 (1989).
4. Schiøtz, J., Di Tolla, F. D. & Jacobsen, K. W. *Nature* **391**, 561–563 (1998).
5. Nieh, T. G. & Wadsworth, J. *Scripta Met. Mater.* **25**, 955–958 (1991).
6. Chen, I.-W. & Rosenflanz, A. *Nature* **389**, 701–704 (1997).
7. Koblinski, P., Phillpot, S. R., Wolf, D. & Gleiter, H. *Interface Sci.* (in the press).
8. Schowalter, L. J. (ed.) *Heteroepitaxy and Strain*, MRS Bull. spec. issue **21**, No. 4 (1996).
9. Tadmor, E. B., Ortiz, M. & Phillips, R. *Phil. Mag.* **A73**, 1529 (1996).
10. Yip, S. (ed.) *J. Computer-Aided Materials Design* **3**, 1–410 (1997).

Transcriptional repression

The cancer—chromatin connection

Ronald A. DePinho

The road leading to cancer has many speed bumps — tumour-suppressor pathways which, when functionally or physically inactivated, result in the cell moving towards a fully transformed state. Mutation of one such pathway, involving the product of the *retinoblastoma* (*Rb*) gene¹, seems to be a rite of passage for almost all cancer cells. The *Rb* protein regulates a key transition in the cell cycle; namely, the commitment to enter the DNA-synthetic (S) phase². In this capacity, *Rb* senses and integrates a multitude of proliferative and anti-proliferative signals, responding to them by regulation of genes that allow progression from a resting G1 state to the S phase².

Cancer biologists have long been preoccupied with understanding how *Rb* acts as a critical downstream effector. Now, reports by Brehm *et al.*³ and Magnaghi-Jaulin *et al.*⁴ on pages 597 and 601 of this issue provide insights into one essential activity of *Rb* — its ability to repress gene expression. These studies have forged a physical and functional link between *Rb* and the histone deacetylase HDAC1, which modulates the architecture of chromatin. Thus, chromatin regulation is implicated in cell-cycle control, and the *Rb*/HDAC1 complex could be a potential target of transforming viruses.

The ability of *Rb* to control the G1/S transition is mediated largely through its interactions with E2F. This protein belongs to the E2F/DRTF1 family, a group of transcriptional activators whose target genes are involved in DNA replication and control of cell growth⁵. On exposure to proliferative sig-

nals, the G1 cyclin-dependent kinases CDK4 and CDK6 are activated, and they phosphorylate the *Rb* component of *Rb*/E2F. At mid-to-late G1, *Rb* becomes hyperphosphorylated and is released from the promoter-bound E2F, allowing transcription of E2F-regulated genes. This marks the cell's irreversible commitment to enter S phase. Thus, *Rb* is a potent repressor when anchored to E2F-regulated promoters. This repression is alleviated by CDK-directed phosphorylation of *Rb*, enabling the expression of genes that are essential for cell-cycle progression (Fig. 1, overleaf).

The *Rb* protein contains a domain (known as the 'A/B pocket') that can silence transcription by sequestering the E2F transactivation domain⁶ and by active transrepression^{1,7} (Fig. 1, top). This active mechanism seems to be the main way in which *Rb* acts, as shown by the ability of chimaeric proteins (fusions between the A/B pocket and a transactivation-domain deletion mutant of E2F) to repress E2F target genes in *Rb*-deficient cells⁸. Brehm *et al.*³ and Magnaghi-Jaulin *et al.*⁴ now provide molecular insights into how *Rb* functions as a powerful transrepressor. They show that E2F, *Rb* and HDAC1 exist as a complex in mammalian cells. Moreover, anti-E2F immunoprecipitates contain HDAC1 protein and histone deacetylase activity, but only in the presence of *Rb*. This result indicates that *Rb* serves as a physical bridge, tethering the activity of HDAC1 to E2F and, by association, to E2F-responsive promoters. These interactions seem to be physiologically relevant to cell-

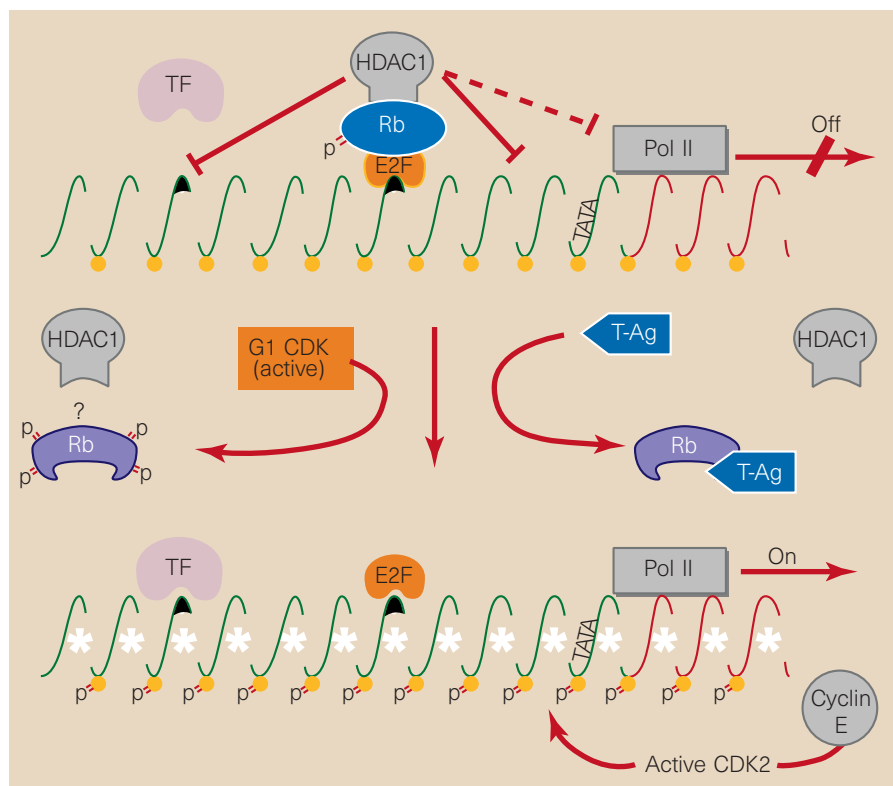


Figure 1 Retinoblastoma and the road to cancer. Retinoblastoma (Rb) is a tumour-suppressor protein that represses gene expression by modulating the architecture of chromatin. Brehm *et al.*³ and Magnaghi-Jaulin *et al.*⁴ have found that Rb tethers the E2F protein to the histone deacetylase HDAC1, to form a complex that prevents expression from E2F-bound promoters. HDAC1 may facilitate the removal of highly charged acetyl groups (white asterisks) from core histones, causing a tighter association between the DNA and nucleosomes, and preventing transcription factors (TF) from gaining access to the DNA. This repression is released when, on exposure to proliferative signals, G1 cyclin-dependent kinases (CDKs) phosphorylate Rb. Viral oncoproteins such as T-Ag may also bind Rb, causing it to dissociate, and allowing transcription to occur from the E2F-bound promoters. (Figure provided by Lynda Chin.)

cycle control because Rb and HDAC1 can act together to repress the cyclin E promoter, and trichostatin A (a potent inhibitor of HDAC1) can alleviate this repression.

The issue of how HDAC1 represses an E2F-regulated promoter remains open. The widely held view is that HDACs mediate the removal of highly charged acetyl groups from core histones. This causes a tighter association between DNA and nucleosomes, thereby impairing the access of transcription factors to DNA-recognition elements⁹. Conversely, histone acetyltransferase-mediated acetylation would destabilize nucleosomes, increasing the accessibility of promoters⁹. Non-histone proteins may also be affected by HDAC1, particularly those involved in the regulation of gene expression. Such speculations are supported by the fact that basal¹⁰ and sequence-specific¹¹ transcription factors are targets of acetylation.

The findings reported by Brehm *et al.*³ and Magnaghi-Jaulin *et al.*⁴ join an expanding number of studies that implicate chromatin modulation in the genesis or suppression of cancer. For example, the E1a oncoprotein can stimulate proliferation by disrupting the growth-suppressive interaction

between p300/CBP and another histone acetyltransferase, P/CAF (ref. 12). A further example of the cancer–chromatin connection comes from observations that the potent anti-oncogenic activities of the Mad(Mx1)/Sin3 complex correlate with its ability to recruit HDAC (ref. 13).

Repression of E2F-bound promoters by Rb is considered to be one of the key mechanisms by which Rb induces growth arrest¹. As such, the interaction between HDAC1 and Rb may represent a main point of attack for the oncogenic process. To test this hypothesis, Brehm *et al.* and Magnaghi-Jaulin *et al.* conducted two types of experiment. One showed that tumour-derived point mutations and deletions that impinge on the A/B pocket abolish both Rb-induced repression and Rb-associated deacetylase activity, correlating with increased transactivation through E2F. The other experiment documented that specific viral oncoprotein sequences can disrupt the interaction between Rb and HDAC1^{3,4}, and displace deacetylase activity from Rb *in vitro*³. This finding is significant because viral oncoproteins invariably target and inactivate the critical components of growth-suppressor



100 YEARS AGO

By utilising the general ideas of the nebular hypothesis of Laplace, and by applying the equations obtained in the first treatise, [Severenus J. Corrigan] proceeds to investigate the genesis and development of the solar system, to determine the ages and temperatures of the planets, as well as a multitude of other important facts, which, if they could only be demonstrated, would place the author on a pedestal by the side of Newton as the greatest astronomer of the age. The fertility of resource of the author in developing his ideas is astonishing, and though at all times the theories are intended to be primarily based on known experimental data, this basis is in many cases so slight and uncertain, and the assumptions so numerous, that the results must be looked upon as mere speculations. The author is equally at home discussing the cause of the Noachian deluge, the nature of vegetation on the planet Mars, and the cause and origin of X-rays.

From *Nature* 3 February 1898.

50 YEARS AGO

Versatility is the most delusive of the fairy gifts; the men of genius on whom it was bestowed otherwise than in subtle malevolence can be counted on the fingers. In the age of Euler himself, Johann Heinrich Lambert shone by his own light, not as a reflexion of the great luminary, and had he consented to be only a mathematician, the course of mathematical history would have been different. But condemned by the humiliations of his early life to demand intellectual submission from everyone he met, Lambert could not bear the thought that there were branches of learning of which he was not a master, or stimuli to which his mind did not respond. Asked by the King of Prussia, "What do you know?" he answered, "Everything, sire"; and to the further question, "How did you learn?", the reply was, "I taught myself". His contemporaries were duly impressed by the range of his knowledge, and if the solipsistic manner of which we are told suggests that he often knew that he was relying on a bluff which might easily be called, his portrait, which we can study for ourselves, suggests that he thoroughly enjoyed the sensation of carrying off the bluff.

From *Nature* 7 February 1948.

pathways such as Rb and p53. The sequences in HDAC1 that are required for contact with Rb also have some homology with the sequence LXCXE. This motif is required for contact with, and functional inactivation of, the Rb protein^{14,15} by the G1 cyclins and viral oncoproteins.

Based on these properties, we can now envisage a sequence of molecular events in which viral oncoproteins clear HDAC1 activity from E2F-responsive promoters (Fig. 1, bottom). The local destabilization of nucleosomes would allow transcription factors to bind neighbouring *cis*-elements and, together with E2F, to stimulate the transcription of critical E2F-regulated genes such as cyclin E. But the chromatin story does not end here. Accumulation of cyclin E results in the activation of CDK2, a favourite substrate of which seems to be the linker histone, H1 (refs 16 and 17). Linker histones are thought to be involved in higher-order packaging of chromatin, and in gene regulation *in vivo*¹⁸. Phosphorylation of H1 enables the chromatin template to open on a global level — a process that is thought to facilitate DNA replication.

As cancer biologists travel the road leading to malignancy, many signposts remind us that the regulatory machinery that governs the architecture of chromatin is an important

barrier to cellular transformation. An improved understanding of the principles that underlie the cell-cycle control of chromatin structure will, therefore, provide opportunities for the design and construction of more effective roadblocks to cancer. □

Ronald A. DePinho is at the Albert Einstein College of Medicine, 1300 Morris Park Avenue, C607 Bronx, New York 10461, USA.

e-mail: depinho@aecom.yu.edu

1. Sellers, W. R. & Kaelin, W. G. Jr *J. Clin. Oncol.* **15**, 3301–3312 (1997).
2. Weinberg, R. A. *Cell* **81**, 323–330 (1995).
3. Brehm, A. *et al. Nature* **391**, 597–601 (1998).
4. Magnaghi-Jaulin, L. *et al. Nature* **391**, 601–605 (1998).
5. Nevins, J. R. *Science* **258**, 424–429 (1992).
6. Flemington, E. K., Speak, S. H. & Kaelin, W. G. Jr *Proc. Natl Acad. Sci. USA* **90**, 6914–6918 (1993).
7. Weintraub, S. J., Prater, C. A. & Dean, D. C. *Nature* **358**, 259–261 (1992).
8. Sellers, W. R., Rodgers, J. W. & Kaelin, W. G. Jr *Proc. Natl Acad. Sci. USA* **92**, 11544–11548 (1995).
9. Wolffe, A. P. *Nature* **387**, 16–17 (1997).
10. Imhof, A. *et al. Curr. Biol.* **7**, 689–692 (1997).
11. Gu, W. & Roeder, R. G. *Cell* **90**, 595–606 (1997).
12. Yang, X.-J. *Nature* **382**, 319–324 (1996).
13. Alland, L. *et al. Nature* **387**, 49–55 (1997).
14. Matusushime, H., Roussel, M. F., Ashmun, R. A. & Sherr, C. J. *Cell* **65**, 701–713 (1991).
15. Figge, J., Webster, T., Smith, T. F. & Paucha, E. *J. Virol.* **62**, 1814–1818 (1988).
16. Matusushime, H. *et al. Mol. Cell. Biol.* **14**, 2066–2076 (1994).
17. Herrera, R. E., Chen, F. & Weinberg, R. A. *Proc. Natl Acad. Sci. USA* **93**, 11510–11515 (1996).
18. Steinbach, O. C., Wolffe, A. P. & Rupp, R. A. W. *Nature* **389**, 395–399 (1997).

carbon. As Hartnett *et al.* say, the implication is that certain classes of organic substrates can be degraded only by aerobic bacteria, and that it is the relative decomposition and preservation of these materials that ultimately limits carbon burial.

My view is that this generalization holds at the high end of the oxygen exposure scale because, in these circumstances, bioturbation and time ensure that labile and resistant organic components (including organic molecules adsorbed to mineral surfaces) are mixed together and exposed to the indiscriminate enzymes and reactive oxygen-containing radicals produced by aerobes⁴. But once oxygen exposure becomes minimal, the definition of exposure time may need to be extended to include the secondary electron acceptors, NO₃⁻, MnO₂, Fe₂O₃, SO₄²⁻ and the organic compounds reduced in fermentation. In other words, a subset of the organic materials that could be oxidized by aerobic respiration is likely to be also metabolizable by anaerobic pathways. If that subset is a significant fraction of the total organic matter falling to the sea floor, and if oxygen exposure is too short to eliminate it, exposure to secondary oxidants must affect preservation.

This line of reasoning may help to explain the great variability in burial efficiency seen in Hartnett and colleagues' data from the Mexican continental slope, one of their sampling areas. This is important because these data anchor their Fig. 2. It also raises various questions. How large is the fraction of sinking particulate carbon that can be degraded by anaerobes? How different are the compounds that can be decomposed by anaerobes from those that are subject to aerobic degradation? And what influence does time or environment have in determining what materials can be degraded either aerobically or anaerobically?

Lastly, there is another ready extension of Hartnett and colleagues' analysis which incorporates situations where oxygen exposure cannot be deduced correctly from the depth oxygen penetrates into the sediment and the apparent sedimentation rate. These are the cases of episodic exposure to oxygen caused by physical events that interrupt sediment accumulation. For example, Aller⁶ has proposed that deltaic sediments have abnormally low rates of carbon burial because the upper 0.5–2 m or so of these muds is intermittently 'fluidized' by waves and coastal currents (that is, caused to flow because of injections of sea water). Fluidization, resuspension and porewater advection may all re-expose buried materials to oxygen.

In light of these processes, one can imagine two sediments that receive the same annual carbon flux and are overlain with the same bottom-water oxygen concentration, but that have very different burial efficiencies. In one, little organic carbon is preserved

Carbon cycle

Feedbacks from the sea floor

Clare E. Reimers

The Earth's sedimentary rocks hold more than 10⁷ gigatonnes of organic carbon of marine origin¹. Scientists see this as evidence that oxygen, produced by photosynthesis, was added to the ocean and the atmosphere in varying degrees, and at the expense of carbon dioxide, over geological time. Scientists also agree that feedback mechanisms have come to limit the accumulation of excess organic carbon in marine sediments worldwide. So what are the feedbacks that operate within the ocean, or between the ocean, land and atmosphere, to limit carbon burial?

On page 572 of this issue², Hartnett *et al.* present correlative evidence that the time organic materials are exposed to dissolved oxygen after incorporation into sediments is one of the crucial controls. What is attractive about this seemingly simple observation is that it unites two variables: carbon 'rain' rate through the water column, and bottom-water oxygen content, each of which has had its advocates as the main influence on the fraction of the total flux of organic carbon to the sea floor that is buried or, conversely, oxidized^{3,4}. A rise in carbon rain rate (or concomitant sedimentation rate³) decreases the oxygen exposure time by depositing greater

amounts of reactive organic matter. This fuels a higher rate of benthic oxygen consumption, which in turn steepens and shoals the profiles of dissolved oxygen concentration in sediment pore water.

In most continental margin and deep-sea depositional settings, profiles of porewater oxygen reflect a balance between oxygen consumption within the sediment, and oxygen transport that results from diffusion and biological activities (irrigation of burrows, for example). Low concentrations of oxygen in bottom waters exclude most groups of large bottom-dwelling organisms and severely limit the diffusive supply. So the depth of oxygen penetration into the sediment also depends on the bottom-water oxygen content^{3,5}.

However, it remains puzzling why oxygen exposure should be critical at all, and why the effects of oxygen exposure on the efficiency of organic carbon burial (as seen in Hartnett and colleagues' Fig. 2 on page 573) are noticeably nonlinear. In sediments, oxidation of organic matter is carried out to a large extent by microorganisms. Many of these are anaerobic, meaning they are consortia of bacteria which use electron acceptors other than oxygen in catabolic reactions that oxidize organic

because physical factors lengthened the effective oxygen-exposure time experienced by the residual organic materials. In the other, buried organics are enriched because what was initially laid down experienced only a single redox cycle.

It will always be difficult to quantify oxygen exposure time in cases where there has been significant physical reworking of sediments. But Hartnett and colleagues' and Aller's recognition of these effects goes a long way towards rationalizing why there are many exceptions to general correlations between carbon burial and individual environmental variables. □

Human genetics

Hair apparent

Kevin Davies

In the United States last year, an eminent television sports commentator was accused of assaulting a female acquaintance in his hotel room. The tawdry testimony highlighted some perplexing aspects of the celebrity's lifestyle, including his alleged penchant for wearing ladies' lingerie, and his trademark toupee, which one witness said she inadvertently yanked off his head during a contretemps. Although sporting a bad hairpiece may have been the least of the commentator's problems, losing one's hair is a common and often distressing event. However, there is hope. Writing in the 30 January issue of *Science*¹, Angela M. Christiano, Wasim Ahmad and colleagues report the first documented genetic mutation associated exclusively with hair loss, or alopecia, in humans — a result that should bolster attempts to devise a safe and effective treatment for baldness.

The most common form of hair loss is androgenetic alopecia — or male pattern baldness — which spreads characteristically from the temples. In alopecia andrea the hair loss is more sudden, occurs in patches, and is frequently attributed to stress. Each year in the United States, an estimated 60 million people (two-thirds of them men) spend an astounding \$1.5 billion to combat baldness. Yet only two treatments have been sanctioned by the US Food and Drug Administration. The first, minoxidil (produced by Upjohn), works in fewer than half the people who use it. The other is finasteride (sold as Propecia by Merck & Co.), which is a reduced dosage of a prostate medication that inhibits production of dihydrotestosterone. The drug is effective but, because it can cause birth defects in women, it is approved only for men. There is, however, a small risk of impotence as a side-effect (which, as one follically challenged writer put it, would be akin to curing a hangnail by lopping off your toe²).

Clare E. Reimers is at the Institute of Marine and Coastal Sciences, Rutgers, The State University of New Jersey, 71 Dudley Road, New Brunswick, New Jersey 08901-8521, USA.

e-mail: reimers@ahab.rutgers.edu

1. MacKenzie, F. T., Ver, L. M., Sabine, C., Lane, M. & Lerman, A. in *Interactions of C, N, P and S Biogeochemical Cycles and Global Change* (eds Wollast, R., MacKenzie, F. T. & Chou, L.) 1–61 (Springer, Berlin, 1993).
2. Hartnett, H. E., Keil, R. G., Hedges, J. I. & Devol, A. H. *Nature* **391**, 572–574 (1998).
3. Emerson, S. in *The Carbon Cycle and Atmospheric CO₂: Natural Variations Archaean to Present* (eds Sundquist, E. T. & Broecker, W. S.) 78–88 (Am. Geophys. Union, Washington DC, 1985).
4. Canfield, D. E. *Chem. Geol.* **114**, 315–329 (1994).
5. Cai, W.-J. & Sayles, F. L. *Mar. Chem.* **52**, 123–131 (1996).
6. Aller, R. C. *Mar. Chem.* (in the press).

Baldness is often termed a 'polygenic' condition, but the truth is that, despite published accounts of hundreds of mendelian hair-growth disorders, we know what only a few of these genes are. One is the gene that is defective in anhidrotic ectodermal dysplasia³, a condition in which sparse hair growth is accompanied by abnormalities in the teeth and sweat glands. These symptoms suggest a defect in some common aspect of epithelial–mesenchymal signalling. Another is a brittle-hair condition called monilethrix, which can be caused by mutations in one of the type II keratins⁴. Even less is known about disorders of excessive hair growth, such as the fascinating X-linked disorder congenital generalized hypertrichosis in which a dense matting of hair covers the face and upper body⁵.

The new study by Ahmad *et al.*¹ concerns a recessively inherited, comprehensive form of baldness called alopecia universalis,



Figure 1 A member of the Pakistani pedigree with alopecia universalis. This man has a reduced number of hair follicles on his scalp, all of which lack hair. (Published with permission.)

manifested in a six-generation Pakistani family descended from the same ancestor. The seven affected members have no body hair, eyelashes or eyebrows (see Fig. 1). To identify the responsible gene mutation, the authors undertook a conventional mapping approach — they typed almost 400 polymorphic DNA markers spaced at regular intervals across the genome, then searched for physically linked clusters transmitted in a tell-tale homozygous pattern in affected people. Such clusters, they reasoned, could well signify the location of the defective gene. Ahmad *et al.* duly found such a region on the short arm of chromosome 8. This region has also been implicated in a linkage search by Nöthen *et al.*⁶, performed on a different Pakistani family with the same condition.

With no obvious candidate genes in sight, Ahmad *et al.* set about finding some of their own. They turned to the mouse, where several excellent examples of abnormal hair development exist. One is the nude mouse, resulting from a frameshift mutation in the *whn* gene⁷ which encodes a winged-helix domain transcription factor. But the authors settled on *hairless*, a pleiotropic mutant strain that is characterized by a lack of hair, impaired thymocyte development and radiation sensitivity. Like the Whn protein, the product of the *hairless* gene is a putative transcription factor — a zinc-finger protein related to the GATA family⁸.

The choice was an inspired one. By cloning the human homologue of this gene, Ahmad *et al.* mapped *HAIRLESS* to the same chromosomal region as the apparent alopecia locus. Not only that, but they discovered that all of the affected people in the Pakistani pedigree harbour two mutated copies of the gene — a missense mutation that is predicted to convert a threonine residue to an alanine near the carboxy terminus. Similar substitutions engineered into the Whn DNA-binding protein are known to abolish its transactivation potential⁹, suggesting that the same may be true of *HAIRLESS*.

The discovery of an alopecia gene is an important event in the history of hirsutism genetics. But its function remains anyone's guess for the time being. Hair growth occurs in unsynchronized cycles¹⁰ consisting of three phases: anagen (growth phase), catagen (shortening phase) and telogen (resting phase). *HAIRLESS* may regulate one of the transitional parts of this pathway. A long list of cytokines and growth factors — including members of the epidermal, fibroblast and transforming growth-factor families¹¹ — has been implicated in the hair growth cycle, providing a variety of potential targets for transcriptional control by *HAIRLESS*. Sorting out the mechanism will naturally be of immense commercial, as well as academic, interest. One company is already advocating

liposomes as an effective gene-delivery vehicle for fatigued hair follicles¹². The phrase 'shampoo and set' could soon take on a whole new meaning. □

Kevin Davies is at the Howard Hughes Medical Institute, 4000 Jones Bridge Road, Chevy Chase, Maryland 20815, USA.

e-mail: daviesk@hhmi.org

1. Ahmad, W. *et al. Science* **279**, 720–724 (1998).

2. Kornheiser, T. *Bald as I Wanna Be* (Villard, New York, 1997).

3. Kere, J. *et al. Nature Genet.* **13**, 409–416 (1996).

4. Winter, H. *et al. Nature Genet.* **16**, 372–374 (1997).

5. Figueroa, L. E. *et al. Nature Genet.* **10**, 202–207 (1995).

6. Nöthen, M. M. *et al. Am. J. Hum. Genet.* (in the press).

7. Nehls, M., Pfeifer, D., Schorpp, M., Hedrich, H. & Boehm, T. *Nature* **372**, 103–107 (1994).

8. Cachon-Gonzalez, M. B. *et al. Proc. Natl Acad. Sci. USA* **91**, 7717–7721 (1994).

9. Schuddenkopf, K., Schorpp, M. & Boehm, T. *Proc. Natl Acad. Sci. USA* **93**, 9661–9666 (1996).

10. Hardy, M. *Trends Genet.* **8**, 55–61 (1992).

11. Danilenko, D. M., Ring, B. D. & Pierce, G. F. *Mol. Med. Today* **2**, 460–467 (1996).

12. Li, L. & Hoffman, R. M. *Nature Med.* **1**, 705–706 (1995).

Materials science

Disciplines bound by pressure

Paul F. McMillan

Materials are the stuff of technology, and the stuff of planets too. Because geologists must study the properties of Earth's materials under the high pressures that prevail inside the planet, they have driven many of the recent advances in high-pressure experimentation. But understanding the properties of materials under extreme compression is also a well-developed branch of condensed-matter physics; and in chemistry, high pressure is used to form new dense crystalline materials, often with unusual coordination and valence states, or interesting electronic and magnetic properties. At a meeting late last year*, these disciplines came together in a Materials Research Society symposium organized around the topic of high-pressure materials. The result was an exciting blend of theory, experiment and speculation from physics, Earth science, chemistry, materials science and engineering.

One of the classic uses of high pressure in materials synthesis has been in the search for new super-hard materials. In the late 1980s, theoretical calculations and empirical systematics led to the prediction¹ that carbon nitride (C₃N₄) with the β-silicon-nitride structure should be even harder than diamond, the hardest known material. Intense efforts followed to synthesize high-pressure forms of carbon nitride, and test their properties. One potential sighting of a new cubic phase was reported in a high-pressure diamond-cell experiment on C₆₀ fullerenes, heated by laser in a nitrogen atmosphere (J. Nguyen, Berkeley). Unfortunately the material lost its crystal structure upon decompression.

A different approach involves synthesis and recovery experiments in a large-volume press, being used to study the kinetics of formation of CN_x materials (A. J. Stevens, Harvard), as we still don't fully understand the carbon–nitrogen phase diagram. So far the nitrogen content produced in these experiments hasn't been high enough, and

super-hard C₃N₄ remains elusive.

Shock compression techniques have been invaluable in geophysics, providing essential data on the behaviour of planetary materials under extreme pressure and temperature, such as the melting of iron in the Earth's core (T. Ahrens, Caltech). Shock processing also has great potential for materials synthesis: new families of dense metastable solids can be expected because of the short timescale of shock-wave synthesis (V. F. Nesterenko, Univ. California; T. Sekine, Tsukuba, Japan). Materials such as hexagonal diamond and tetragonal BN should have interesting electronic properties. Some of the most exciting new developments are in laser shock processing, which produce the most extreme pressure–temperature conditions currently available, useful for studying molecular dissociation and ionization of gases into the

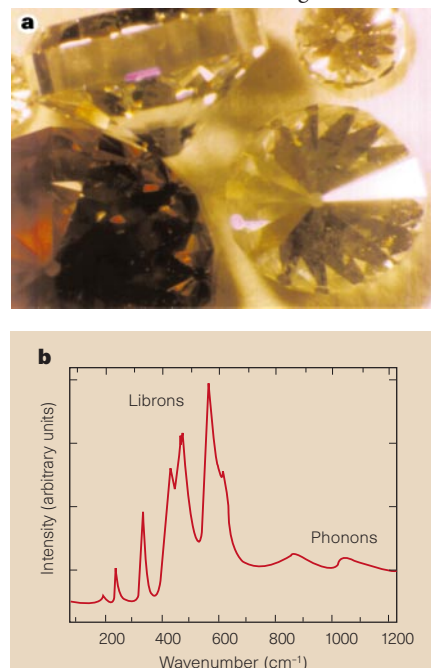


Figure 1 Diamond anvils. Ultra-pure, single-crystal diamonds can now be grown (a), and have already been used as windows to obtain spectra (b) of *para*-hydrogen at multimegabar pressures².

plasma regime (R. Cauble, Lawrence Livermore Natl Lab. (LLNL)). A high-energy neodymium–glass laser has just been developed for industrial shock processing of metals, with penetration of the shock front (1–3 GPa) to deeper than 1 mm (L. Hackel, C. B. Dane, LLNL).

Another focus of excitement in high-pressure physics is the unexpectedly complex behaviour of the simplest material, hydrogen (N. W. Ashcroft, A. L. Ruoff, Cornell; J. Kohanoff, ICTP/SISSA, Trieste). The observation of hydrogen metallization in the high-temperature *fluid* state at about 140 GPa and 3,000 °C (N. C. Holmes, W. J. Nellis, LLNL) is now supported by shock experiments (V. E. Fortov, Inst. Chem. Phys., Chernogolovka). But to experiment on solids at these pressures, the tool one needs is a diamond anvil cell. Decades ago, chemical vapour deposition technology led to the low-pressure synthesis of metastable diamond, and this technique is now being adapted to grow 'designer' diamond anvils with specified chemical or isotopic purity, for a whole new class of high-pressure experiments (Figs 1 and 2) (A. Israel, Y. Vohra, Univ. Alabama, Birmingham). In recent experiments on hydrogen in the solid state, these diamond anvils provide an astonishingly clear spectroscopic view (Fig. 1), allowing entirely new vibrational excitations to be observed² to pressures above 200 GPa. These anvils can also be grown with embedded electrical leads, allowing new types of electrical and magnetic measurements at high pressure and temperature.

Measuring the magnetic properties of

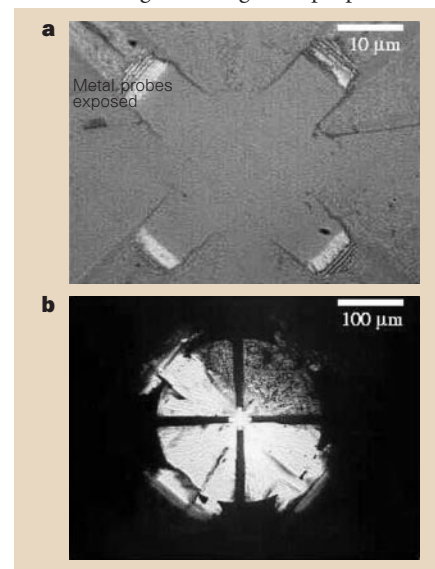


Figure 2 A 'designer' diamond anvil (Y. Vohra, S. Catledge, Univ. Alabama at Birmingham; S. Weir, J. Akella, LLNL). Metallic probes are exposed near the centre of the diamond anvil for making contact with the sample (a). An overall view of the diamond anvil shows encapsulated metal probes with a 10-μm layer of synthetic diamond for electrical insulation (b).

* Fall Meeting of the Materials Research Society, Boston, 1–5 December 1997.

materials at high pressure is challenging, because intrinsic magnetism of the high-pressure cell must not interfere with the measurement (S. Tozer, Natl High Mag. Field Facility, Florida). But using a new technique for measuring magnetic susceptibility that involves immersing the high-pressure cell with its loaded sample in an external magnetic field (V. Struzhkin, Carnegie Inst., Washington), elemental sulphur has been shown³ to become superconducting at high pressure, and to reach T_c values of 17 K at 160 GPa — the highest recorded superconducting transition temperature for an elemental material.

Another new tool is being brought to bear on high-pressure materials — intense beams of X-rays from new-generation synchrotrons (G. Shen, Y. Wang, GSECARS, Univ. Chicago). Beryllium gaskets can be used to permit the X-ray beam to enter a diamond anvil cell from the side⁴ (H.-K. Mao, Geophysical Laboratory, Washington). This means that X-ray absorption experiments on light elements — down to chlorine, sulphur, and perhaps even silicon — can be carried out *in situ*, and also allows a full determination of elastic properties of materials at high pressure.

In the field of optoelectronics, one of the most exciting recent developments is the use of GaN to make blue LEDs and lasers. Growth of large, single GaN crystals (as substrates for homoepitaxy of unstrained nitride films) is experimentally challenging, because the material decomposes far below its melting point at ambient pressure — but high pressure stabilizes GaN enough for large, single crystals of up to 1 cm² surface area to be obtained, with fine control over

impurity content and structural defects (S. Porowski, Warsaw).

Materials science is a field in constant flux, in the search for new technological materials and processes, and ways to probe material properties and structure. The high-pressure route offers another window through which to view materials science, with a largely unexplored landscape waiting beyond. □

Telemicroscopy

Net progress

Kevin Burton and Daniel L. Farkas

The microscope is a staple of the sciences, but for many years only a single observer could effectively use the instrument. Increasing the number of eyepieces and, much more recently, adding cameras and monitors for local or remote viewing, were big advances. But there has remained the need to enable groups physically separated from the microscope to control it interactively, and view the images together.

On page 613 of this issue¹, Wolf *et al.* describe a telemicroscopy system that allows Web-browser software to connect multiple 'clients' through the Internet to a 'server' computer that controls an automated light microscope. Such Internet-based remote control is not new²⁻⁴. But Wolf *et al.* have made telemicroscopy more accessible by employing commonly used Web browsers as remote viewers (see also ref. 3), while taking the approach a step further by incorporating Java technology to make the interface more interactive and dynamic.

Paul F. McMillan is in the Department of Chemistry, Arizona State University, Tempe, Arizona 85282-1604, USA.

1. Liu, A. Y. & Cohen, M. L. *Science* **245**, 841–842 (1989).
2. Goncharov, A. F., Hemley, R. J., Mao, H. K. & Shu, J. *Phys. Rev. Lett.* **80**, 101–104 (1998).
3. Struzhkin, V. V., Hemley, R. J., Mao, H. K. & Timofeev, Y. A. *Nature* **390**, 382–384 (1997).
4. Singh, A. K., Mao, H. K., Shu, J. & Hemley, R. J. *Phys. Rev. Lett.* (in the press).

Several developments lie behind this and other advances in telemicroscopy. Video microscopy and digital imaging allowed acquisition of images for display to large groups on video monitors, as well as electronic transmission to many different sites⁵. Automation of microscopes allowed digital (including remote) control of motorized moving parts, electro-optics and other components, and, along with software control, facilitated more rapid, reproducible and complicated operations⁶.

The final step in freeing the operator from the instrument has been to control the microscope remotely by sending instructions and the resulting images over the Internet. Although remote control of telescopes began some years ago⁷, connection of microscopes to the Internet came later, with the word 'telemicroscopy' first being applied to electron microscopes⁸. Given the cost of electron microscopes, control from remote sites is extremely valuable for both research and education⁹.

As Wolf *et al.* show, there are equal and broader benefits to be had with light microscopy linked to the Internet. They have connected a PC running under Windows to an automated microscope equipped with a colour CCD camera and frame-grabber. The PC acts as a Web server, accepting commands from clients using common Web-browser software and passing the instructions to the microscope and camera through commercial software drivers (Fig. 1). The client can control automated features of an upright bright-field microscope, including stage position, focus, lamp intensity and magnification (by rotating a motorized objective turret). Proper illumination is maintained when changing objectives by automatically adjusting lamp intensity, condenser lens, and field and aperture diaphragms. The software offers some features for image handling that are useful even with microscopes under manual control — for example, detection of changes in the digitized images resulting from local operations, such as adjusting focus or stage position, and then automatic updating for clients.

Like all other systems that use the Internet, this one slows down when transferring

Exhibition

Natural attraction

Later this month, the Natural History Museum, London, presents a little-known aspect of its collections to the public — a selection of 70 works of natural-history art from its massive collection of some half a million. The exhibition is called 'Images from Nature', and birds, bugs and botanical specimens are all predictably well represented. Yet there are also less obvious subjects, such as 'pebbles found on British beaches' and copper ores, as well as a couple of vivid lichens executed in watercolour and pencil.

Shown here is Bryan Kneale's depiction in chalk of a giant tortoise (1990), one of the 'bone drawings' he produced while associated with the museum. Other exhibits come from the great days of exploration and the hands of such illustrators as Ferdinand Bauer and Sydney Parkinson, who sailed with Cook on the *Endeavour* (and died on the return leg of



the journey). The organizers have also failed to resist the temptation to include examples of the design work of the Natural History Museum's architect, Alfred Waterhouse — justifiably so, for his chunky drawings are a refreshing contrast to the delicacy of some of the other pictures.

The exhibition is at Christie's, King Street, London SW1 (not the museum itself), and runs from 9 to 20 February.

Tim Lincoln

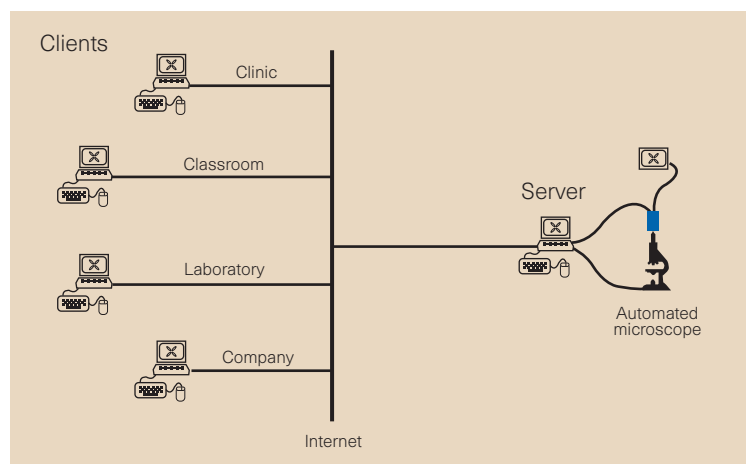


Figure 1
Microscopic connections — an Internet-based system of remote viewing and control as implemented by Wolf *et al.*¹.

large images. Delays result from both the Internet transfer itself, and the process of image compression that precedes transmission^{1,2}. Wolf *et al.* speed things up by providing options for reduced-size images, autofocus to eliminate the transfer of multiple images otherwise required during manual focusing^{2,3,8}, and generally reducing Internet traffic between client and server using Java applets running on the client machine.

Another advantage of Internet-based systems is that many clients can simultaneously view, and confer over, a single specimen^{2-4,8}; the use of common Web-browser software increases this potential enormously^{1,3}. To aid communication between clients, Wolf *et al.* provide a neat Java-based interactive pointer overlaying the image; movements of the pointer by any one client can be seen by all. Further developments would be a facility to annotate images with text or drawings, and provision of video/audio conferencing^{3,4}.

One of the disadvantages of not being at the microscope can be difficulty in getting, and staying, orientated — that is, keeping track of where the current image lies within the entire specimen. Solutions include acquiring a reference image at low magnification, graphically indicating which portion of the specimen is currently being viewed, or building up a montage of contiguous high-magnification images^{3,4,8}.

The uses of telemicroscopy span research, education, medicine and manufacturing^{1-4,8,9}. For centralized research facilities, distant collaborations will benefit by allowing collaborators to stay put yet look in on their specimens during an experiment⁸. Likewise, travelling professors could check in on experiments being carried out at their home laboratory.

Telepathology is a further application^{1,3,4}, consultation around case specimens being aided by moving around the specimen, changing magnification and otherwise examining it for diagnostic features (although image resolution and time response are limitations). In education, large numbers of students or colleagues, on-site or farther afield, could view precious specimens without having to crowd around microscopes or video

monitors. The separation between server and client can also take the form of a protective barrier, for example when high-power lasers are used for illumination, or clean-rooms are required for manufacturing processes or for the examination of infectious specimens². Many servers could also be accessed by a single client, allowing modifications in server software or hardware to be tested from a single site. Finally, other types of instrument are being connected to the Internet, atomic force microscopes being a recent example¹⁰.

Assembling a telemicroscopy system, however, is not yet as easy as downloading the latest Web browser. One consideration is the expense of automated hardware. Another is the compatibility of Java applets with the Web-browser/operating system/computer-platform of the client. Also, more and more systems incorporate some level of telemicroscopy^{1-4,8,9}, but each has distinct features and accessibility that are often affected by regulatory and intellectual property issues. So anyone considering setting up a server has to weigh the various options. Nonetheless, these are still early days. Given the ubiquitous use of microscopes, it seems inevitable that many will become linked into an interactive network of researchers, educators, clinicians and companies through the World Wide Web. □

Kevin Burton and Daniel L. Farkas are in the Center for Light Microscope Imaging and Biotechnology, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA.
e-mail: Burton@a.cfr.cmu.edu

1. Wolf, G., Petersen, D., Dietel, M. & Petersen, I. *Nature* **391**, 613–614 (1998).
2. Kao, J., Troxel, D. E. & Kittipiyakul, S. *Proc. SPIE* **2901**, 90–100 (1997).
3. Maturro, R., Kath, G., Zeigler, R. & Meechan, P. *Biotechniques* **22**, 1154–1157 (1997).
4. Telemicroscopy Systems <http://www.telemicroscopy.org>
5. Inoue, S. & Spring, K. R. *Video Microscopy: The Fundamentals* 2nd edn (Plenum, New York, 1997).
6. Taylor, D. L. *et al. Ann. NY Acad. Sci.* **820**, 208–228 (1997).
7. Automated Telescopes <http://www.telescope.org/rti/automated>
8. Fan, G. Y., Mercurio, P. J., Young, S. J. & Ellisman, M. H. *Ultramicroscopy* **52**, 499–503 (1993).
9. Voelkl, E., Allard, L. F., Bruley, J. & Williams, D. B. *J. Microsc.* **187**, 139–142 (1997).
10. Baur, C. *et al. J. Vac. Sci. Technol. B. Microelectron. Nanometer Struct.* **15**, 1577–1580 (1997).

Daedalus

The seeing ear

Sunlight makes us cheerful. The pineal gland in the brain is responsible; in bright light it releases hormones such as melatonin more strongly. Melatonin is a mood enhancer and sexual stimulant. One consequence is seasonal affective disorder, or SAD, whose sufferers are peculiarly depressed and unsexy during the winter. The short, dull days fail to stimulate their pineal glands. Brief daily exposure to intense light reduces the SADness.

In this connection, Daedalus recalls the theory that the pineal gland is the relic of a third eye possessed by our distant reptilian ancestors. Presumably it still retains a vestigial sensitivity to light. Indeed, it must be amazingly light-sensitive. It is stuck almost in the middle of the brain — how can light reach it? Through the ears, says Daedalus. The pineal gland is almost exactly on the right level, and the ear hole is an excellent entry point for light. It is formed of soft, translucent tissue, and is close to the carotid canal and several nerve pathways into the brain. Light could diffuse along these paths quite well, without being blocked by opaque bone.

This notion has many implications. It explains the sexual allure of the world's tropics and sunlit regions, despite the dark glasses which are *de rigueur* for their devotees. It explains the ear-concealing headgear of monks and subservient groups in some traditional societies: the garb must serve to damp dangerous feelings, and encourage a grave and humble mien. It also suggests a simple method of countering SAD and depression. Daedalus's 'Earlight' is a headband carrying two well-focused lamps to beam light into the wearer's ears. Human tissue is most translucent to red light, so small red lasers or LEDs are the lamps of choice. Accurately aimed, they will illuminate the pineal gland far more effectively than the fiercest daylight. The Earlight should bring cheer and erotic interest to the saddest of SAD sufferers.

More cunning still, the Earlight could illuminate one ear more than the other, thus biasing pineal output into the more illuminated hemisphere. Left-brain-dominated rationalists could boost their emotional awareness, and right-brained dreamers could get a grip on reality. And combined with a personal stereo system, the Earlight could cure the glum truculence of so many teenagers, whose melatonin level is clearly lowered by their permanent wearing of ear-blocking audio equipment.

David Jones

Obituary

Vladimir Prelog (1906–98)

Pioneer of stereochemistry

Until quite recently, the nonagenarian Vladimir Prelog was a familiar figure in the chemistry building of the ETH. Conspicuous in his immaculate, white laboratory coat, he could be seen in the corridors or in the Chemiebar, where he drank his coffee most mornings and afternoons. New students used to ask: "Who is that elderly gentleman in the white lab coat? Is he a laboratory technician?" On being told that this was Prelog, the Nobel prizewinner, they were impressed, for they knew his name from the first-year lectures. We shall see him no more. Vladimir Prelog, one of the great organic chemists of our century, died on 7 January after a short illness. He is survived by his wife and son.

A child of the old Austro-Hungarian empire, Prelog was born in Sarajevo on 23 July 1906. He was present at the fateful visit of the Austrian crown prince, Archduke Franz Ferdinand, on 28 June 1914, when, waiting to throw flowers in front of the state carriage, he heard the shots that presaged the end of that era. His school years were passed in Zagreb, after which he studied chemistry at the Prague Institute of Technology, obtaining his doctorate in 1929 with Emil Votoček but being influenced more by Rudolf Lukeš, who led him into the chemical world of alkaloids. It was the time of the great depression, and Prelog supported himself 'semi-legally' by preparing chemicals in a commercial laboratory whose owner and director became Prelog's first doctoral student. In 1935 he moved back to Zagreb, where he taught at the university, first as lecturer and then as professor. Meanwhile, he was associated with a fledgling chemical company in Zagreb (which later developed into Pliva, the largest Yugoslav pharmaceutical company). His commercial preparation of sulphanilamide (Streptazol) brought the financial support needed for his research.

In 1941, after the German occupation of Yugoslavia, he managed to escape under adventurous circumstances to Zürich in neutral Switzerland. There, at the organic chemistry laboratory of the Federal Institute of Technology (ETH) under his fellow countryman Leopold Ruzicka, he found a temporary refuge, which gradually developed into a home in every sense of the word. If the ETH gave Prelog a home, Prelog brought fame to the ETH, with which he was associated for more than 50 years. After Ruzicka's retirement in 1957,



Prelog became head of the laboratory but soon decided to share his powers, and his responsibilities, with his younger colleagues — this was then in continental Europe a mildly revolutionary step for the director of a chemistry institute. After his retirement in 1976 Prelog promptly registered as a special student (Fachhörer) in order to continue his scientific work. At the ETH this is not regarded as an automatic perquisite of a retired professor; it is a privilege for which one may be grateful. Prelog's 1991 autobiographical sketch was entitled, "My 132 Semesters of Chemistry Studies". Renowned as teacher and researcher, he regarded himself as a student to the end.

As an organic chemist, Prelog bridged the gap between the classical tradition and the modern school. When he began his long research career, molecular structure was still being inferred by classical methods, that is to say, by subtle but indirect arguments from chemical reactivity. Success in this area required profound knowledge of chemistry, rigorous logic and, sometimes, a flight of inspiration as well. By mid-century, physical methods — molecular spectroscopy, X-ray analysis and, later, nuclear magnetic resonance — were beginning to displace the traditional chemical methods. Prelog's early work on the structure of alkaloids from *Cinchona*, *Strychnos* and other sources was done by classical methods, whereas his later studies of metabolic products of microorganisms were largely based on the newer physical methods. It was mainly through his influence that organic chemistry at the ETH was swift to take advantage of the new techniques.

For his contributions to stereochemistry, Prelog shared the 1975 Nobel prize in chemistry with J. W. Cornforth. Stereochemistry is not so much a branch of chemistry as a way of looking at the science, an explicit concern with molecular shape. Prelog was always

fascinated by the spatial forms of organic molecules. One of his early feats in Zagreb was the synthesis of adamantane, with its highly symmetrical molecular structure. Later, in Zürich, he was one of the first to see the importance of steric effects on the reactivity of molecules with medium and large ring systems — for example, the non-reactivity of cyclodecanone towards hydrogen cyanide compared with the ready formation of cyanhydrins of smaller and larger rings. This led him to conformational analysis, and indeed, together with D. H. R. Barton, he was one of the founders of this important branch of modern chemistry. A feature of his later work was the emphasis on steric bulk as a factor in controlling the course of asymmetric synthesis and as a basis for understanding stereospecific reactions in general.

In the mid-1950s, these interests led Prelog to develop, together with his British colleagues R. S. Cahn (then editor of *The Chemical Society*) and C. K. Ingold, the now universally adopted 'CIP' system of rules for specifying the sense of chirality at tetrahedral stereogenic centres. Towards the end of his life, Prelog was preoccupied with the complexities inherent in this system. There can be few organic chemists of his generation who have had the intellectual strength to master the more mathematical, graph-theoretical aspects of the models they use to describe molecular structures and their interconversions.

Not only as a chemist but also as a person, Prelog made a powerful impression on others. He was admired as a scientist and loved as a person by a multitude of friends from all over the world. A formidable lecturer and talker in general, renowned for his inexhaustible supply of jokes, Yugoslavian folk-sayings, and anecdotes about every famous chemist who has ever lived, he captivated all. Ambitious as far as science was concerned, he was uninterested in the exercise of power — he said that he disliked making decisions that affected other people. He was seldom provoked to anger and claimed to have an unusually low adrenaline level. But, of course, behind this outward appearance there was a serious, introspective side. His ironic and often self-deprecatory humour was frequently a protective wall for his inner self. Sadly, his last years were clouded by the tragic course of events in his native land, which caused him immeasurable grief and sadness.

Jack D. Dunitz

Jack D. Dunitz is in the Organic Chemistry Laboratory, Swiss Federal Institute of Technology, ETH-Zentrum, CH-8092, Zürich, Switzerland. e-mail: dunitz@org.chem.ethz.ch

Onwin's holistics

Crystals forming on a sea of brine in one of Glen Onwin's installations demonstrated the emergence of shape from chaos – and reflected the artist's interest in the origins of matter.

Martin Kemp

Early in the 1970s the artist Glen Onwin discovered the extensive saltmarsh near Dunbar to the east of Edinburgh. Its landscape is shaped by tidal flooding, the ebb and flow of subterranean waters, erosion and consolidation, precipitation and evaporation, the crystallizing and dissolving of salt, and the tenacious grip of specialized vegetation. This strange topography embodied through natural processes those cyclical qualities of life and death that Onwin had already been seeking through artistic creation.

Onwin works with a wide range of media: photographs, not only of features visible to the naked eye but also of structures discernible with an electron microscope; printed images, including a large white-on-black recitation of the periodic table of the elements; semi-abstract pictures 'painted' with wax, minerals and organic materials, composed at least in part by chance; and installations that typically use concentrated solutions of chemicals to create evolving panoramas of seas, lakes, continents and islands.

His 1992 installation in the partly derelict Square Chapel in Halifax, West Yorkshire, centred upon a vat of black brine, a square sea in which clusters of crystals congealed around barren lands of floating wax, each different in their aggregated morphologies yet essentially similar in the underlying mechanisms of aggregation.

The processes behind such formations stand within the province of the science of chemistry. And there is much in Onwin's response to the phenomena that is consistent with modern science, in particular the search for the constants that decree that salt ideally seeks to crystallize in cubes, or that determine those common morphologies transcending scale, from the cosmic to the microscopic.

He says: "I want to work in a microcosmic, macrocosmic way, to get as much information from the larger image of the work as... from the detail. It is important you know the detail is there, but you should stand back... and view the whole work with that knowledge."

However, he is also delving into realms of symbolism that share more with alchemy, particularly in its psychological interpretation by Carl Jung, than with the prevailing tenor of modern science. Such formations as a cube of salt, and a circular ring of vegetation growing from its dying centre in the saltmarsh, echo the ancient symbols that express the mystical



Onwin's *Nigredo* — *Laid to Waste*, 1992 (above), installation in the Square Chapel, Halifax. Saltmarsh, Dunbar, 1974 (right).

quest to explain age-old perceptions. His fascination with the emergence of shape from chaos marries the *massa confusa* of alchemy with the modern speculations on the origins of matter. If this seems a strange marriage, it is appropriate to recall that Newton was an alchemist. The title of the Halifax installation, *Nigredo*, is openly alchemical in its reference to the primitive and sometimes recurrent stage of black putrefaction in the transmutation of substances, while its subtitle, *Laid to Waste*, alludes to his long-standing concern for the fragile ecology of our planet.

In the saltmarsh, a precarious balance



determines that the same substances can either give life or decree death. He senses that this stands as both physical model and imaginative metaphor for the holistic processes, small and grand, in which we intervene at our peril. □
Martin Kemp is in the Department of the History of Art, University of Oxford, 35 Beaumont Street, Oxford OX1 2PG, UK.

COURTESY OF HENRY MOORE SCULPTURE TRUST AND GLEN ONWIN

GLEN ONWIN

Worms bask in extreme temperatures

Temperature is one of the most important environmental factors that governs a species' distribution. Some highly specialized prokaryotes can grow at temperatures above 113 °C (ref. 1), but eukaryotes appear less versatile² and do not normally occur above 55 °C. Here we show that a colony-dwelling polychaete worm, inhabiting deep-sea hydrothermal vent chimneys, regularly experiences temperatures above 80 °C and a thermal gradient of 60 °C or more over its body length.

Active deep-sea hydrothermal vents typically contain areas of high temperature flow (above 300 °C) which are essentially abiotic. Associated with and peripheral to these hot sites are cooler (below 100 °C) diffuse flow regions where mixing with ambient seawater (2 °C) occurs. The thin tube dwellings of the Pompeii worm (*Alvinella pompejana*; Fig. 1) form thick, heavily channelled structures along the outer walls of vent 'chimneys' created by an accumulation of metal sulphides. Although these worms spend most of their time in the tubes with their gills and buccal structure extended from the tube opening, individuals make regular excursions of up to 1 m to feed on free-living filamentous bacteria growing on the outer surface of the colony.

We deployed self-recording time-lapse temperature probes from the research submersible DSV *Alvin* at high-temperature chimneys located at the M-Vent site (9° 50.6' N, 104° 17' W) within the Axial Summit Caldera on the East Pacific Rise (EPR) during November 1995 and April 1996. M-Vent is one of three active chimneys, 5–7 m tall, from which intense diffuse vent fluids emerge through densely populated Pompeii worm colonies.

On each dive we positioned a specially designed time-lapse temperature logger (the 'Mosquito'), containing a 20-cm-long titanium probe (0.75 cm diameter), with its temperature sensor adjacent to a Pompeii worm's tail (Fig. 2a). We placed two further temperature recorders adjacent to the tube of interest. We chose completely intact worm tubes, to avoid temperature sampling where the hydrodynamic conditions of the colony might have been altered; we also monitored unoccupied tubes. In all cases resident worms displayed normal behaviour before, during and after deployment of the temperature recorders. The loggers were calibrated before and after each dive. We also made real-time temperature measurements with *Alvin*'s low-temperature probe (0–60 °C range).

In six independent surveys, temperatures recorded 6–8 cm within the worm tubes averaged 68 ± 6.33 °C (mean ± s.d.),



Figure 1 *Alvinella pompejana*, the Pompeii worm, found colonizing the sides of active deep-sea hydrothermal vents (actual length 6 cm). The fleece-like covering on the dorsal surface of the worm is composed of a complex epibiotic microbial community dominated by filamentous bacteria.

with frequent spikes exceeding 81 °C (Fig. 2b). The external temperatures were variable, depending on the placement of the *in situ* recorders and the submersible's low-temperature probe. Measurements made at the tube openings averaged 22 ± 2.5 °C; thus, a temperature gradient of up to 60 °C exists along the length of the worm's body. These measurements were consistent with long- and short-term measurements taken near Pompeii worm colonies at several different EPR sites^{3,4}. No temperature differences were detected between occupied and unoccupied tubes.

A Pompeii worm has survived short exposure to 105 °C (ref. 5); however, this was outside the worms' natural habitat. Our data demonstrate unambiguously that Pompeii worms inhabit an environment with an unprecedented temperature gradient of up to 60 °C. This gradient, although somewhat dynamic, is remarkably uniform over a period of hours. Such a consistent and pronounced gradient over the length of a single worm establishes *A. pompejana* as the most eurythermal (tolerant of wide temperature range) organism on record. Moreover, the 81 °C temperature at the posterior of the worm suggests that this is also the most thermotolerant metazoan known.

The most thermotolerant and eurythermal eukaryote previously described is the Sahara Desert ant *Cataglyphis*, which forages under the midday sun at temperatures nearing 55 °C and remains under-

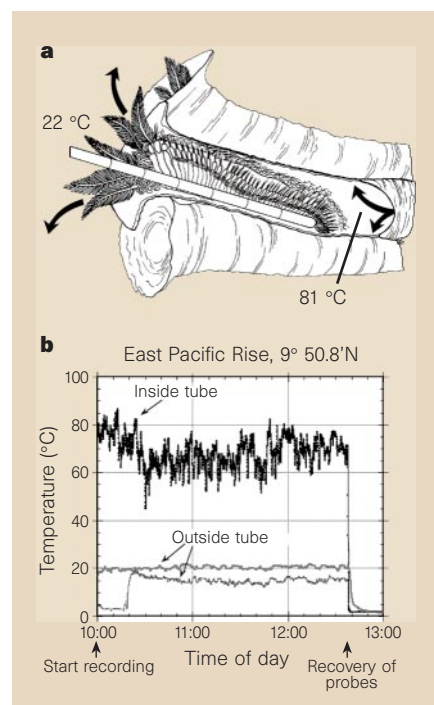


Figure 2 Temperature records of Pompeii worm habitat. **a**, A cut-away representation of the temperature probe placement in a Pompeii worm tube. Temperatures are maximal values obtained during one deployment. **b**, Inside and two representative outside temperature profiles over a 3-h period for a single deployment. The temperature recorders made measurements at 2-s intervals. The micro-variation observed in the profiles was probably generated by the turbulent mixing of fluids throughout the colony. Spectral analysis of the temperature records provided no evidence of tidally influenced cycles or strong harmonic periodicities.

ground for the rest of the day while the temperature rarely drops below 20 °C (ref. 6). These conditions are not as extreme as those experienced by the Pompeii worm.

Several studies have proposed that the Pompeii worm tube might be an efficient barrier against the thermal effect of the diffusing fluid^{3,7}. Other researchers have proposed active ventilation or convective cooling as mechanisms to minimize the thermal exposure of the worm^{3,8}. The present study suggests that the tubes actually channel the hot diffuse flow fluids over the worm's body, maintaining a continually flushed high-temperature environment.

How the Pompeii worm's physiology has adapted to function under such an extreme temperature gradient has yet to be determined. It is also unclear how a metazoan's cytoskeletal proteins, chromosomes, nucleic acid processing machinery and other macromolecules, previously thought to be the thermally limiting components of multicellular eukaryotes, can function at temper-

atures above 80 °C. Recent interest in the Pompeii worm has focused on a unique assemblage of symbiotic filamentous proteobacteria that cover the dorsal surface of the animal^{9,10}. Like their hosts, these bacteria are positioned within this extreme thermal gradient and survive the same high-temperature environment laden with heavy metals and hydrogen sulphide. Studies of the worm and its associated microflora afford a unique opportunity to discover the biochemical adaptations that allow organisms to thrive in such an extreme thermal regime.

S. C. Cary

College of Marine Studies, University of Delaware,
Lewes, Delaware 19958, USA

e-mail: caryc@udel.edu

T. Shank

Institute of Marine and Coastal Sciences, Rutgers,
State University of New Jersey, P.O. Box 231,
New Brunswick, New Jersey 08903-0231, USA

J. Stein

Diversa Corporation, 10665 Sorrento Valley Road,
San Diego, California 92121-1623, USA

1. Blochl, E. et al. *Extremophiles* **1**, 14–21 (1997).
2. Tansey, M. R. & Brock, T. D. *Proc. Natl Acad. Sci. USA* **69**, 2426–2428 (1972).
3. Chevaldonné, P., Desbruyères, D. & Le Hêtre, M. *Deep Sea Res.* **38**, 1417–1430 (1991).
4. Desbruyères, D. et al. *C. R. Acad. Sci. Paris* **295**, 489–494 (1982).
5. Chevaldonné, P., Desbruyères, D. & Childress, J. J. *Nature* **359**, 593–594 (1992).
6. Gehring, W. J. & Wehner, R. *Proc. Natl Acad. Sci. USA* **92**, 2994–2998 (1995).
7. Gaill, F. & Hunt, S. *Mar. Ecol. Prog. Ser.* **34**, 267–274 (1986).
8. Desbruyères, D., Gaill, F., Laubier, L. & Fouquet, Y. *Bull. Biol. Soc. Wash.* **6**, 103–116 (1985).
9. Haddad, M. A., Camacho, F., Durand, P. & Cary, S. C. *Appl. Environ. Microbiol.* **61**, 1679–1687 (1995).
10. Cary, S. C., Cottrell, M. T., Stein, J. L., Camacho, F. & Desbruyères, D. *Appl. Environ. Microbiol.* **63**, 1124–1130 (1997).

Phytoplankton change in the North Atlantic

A marked increase in global temperature over the last century was confirmed by the second Assessment Report of the Intergovernmental Panel on Climate Change¹. Here we report significant positive and negative linear trends from 1948 to 1995 in phytoplankton measured by the Continuous Plankton Recorder survey in the northeast Atlantic and North Sea that might reflect a response to changing climate on a timescale of decades. Spreading of unusually cold waters² from the Arctic might have contributed to the decline in phytoplankton north of 59° N. Further south, phytoplankton season length and abundance seem to have increased.

The Continuous Plankton Recorder (CPR) survey³ is the only biological monitoring programme that operates on an ocean basin scale. Merchant ships voluntarily tow CPRs, at a depth of ~10 m, on their

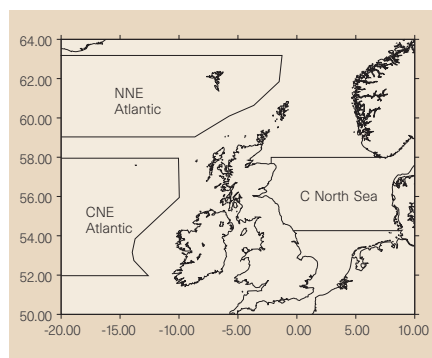


Figure 1 A map defining the boundaries of the three areas for which Phytoplankton Colour data have been averaged. The eastern boundary of the Atlantic areas coincides approximately with the 200-m contour at the shelf edge.

normal routes of passage⁴. The survey is unique in that the methodology and plankton analysis procedures for ~400 taxa and Phytoplankton Colour have been maintained with little change for more than 50 years.

Phytoplankton Colour is a visual index of chlorophyll based on the intensity of the green coloration of the CPR filtering silk, which is assigned numerical values in four categories. Calculated annual primary production averaged for seven areas of the North Sea⁵ is statistically comparable to mapped Phytoplankton Colour averaged for the same areas ($r^2=0.9$, $p<0.01$), suggesting that the colour index might also reflect variation in plant production in the sea.

Results averaged for each month during 1948–95 for three areas of the North Sea and northeast Atlantic (Fig. 1) were analysed. An increasing trend is evident in the North Sea and the Atlantic between 52° and 58° N (Fig. 2), with evidence for a step-wise increase after the mid-1980s. The results for these areas show similar changes in trend and increased season length over the years 1981–91 to a satellite-derived vegetation index for the Northern Hemisphere above 50° N (ref. 6). An inverse pattern of change is seen in Phytoplankton Colour for the northern northeast Atlantic in an area that has tended to experience colder surface temperatures.

The eastward spread of negative-anomaly sea-surface temperatures towards the Faeroes in the northern northeast Atlantic² is most likely to be related to progressive changes in regional atmospheric forcing. Changes in the North Atlantic Oscillation have altered the centre of deep-water convection from the Greenland Sea to the Labrador Sea after 1988 (ref. 7). The likely sources of these colder waters are stronger westerlies, increased formation of Arctic surface water in the Greenland Sea and/or a larger export of fresh water from melted ice⁸ and permafrost⁹ in and around the Arctic Sea as a response to high

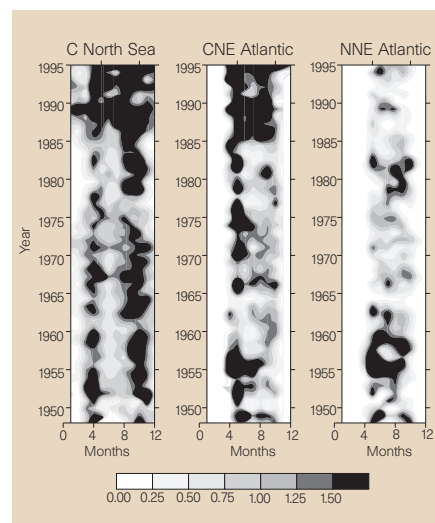


Figure 2 Contour plots of mean monthly Phytoplankton Colour during 1948–95 for the central (C) North Sea, central northeast (CNE) Atlantic and northern northeast (NNE) Atlantic. A regression applied to the data with the seasonal signal removed by a centred 12-month running mean is highly significant: NNE Atlantic $r^2=-0.46$, CNE Atlantic $r^2=0.59$, central North Sea $r^2=0.52$; $p<0.001$ in all three cases.

positive-temperature anomalies in northern Eurasia and Alaska.

Our observations provide the first marine surface time-series evidence for a vegetation response to what seems to be climatic forcing. This has implications for CO₂ fluxes and the productivity of the North Atlantic. The results draw attention to the possible importance of shelf phytoplankton growth to the global carbon budget, and highlight the paucity of comparable information that exists as time series for other shelf regions and for the Southern Hemisphere.

Philip C. Reid, Martin Edwards*, Harold G. Hunt, Andrew J. Warner

Sir Alister Hardy Foundation for Ocean Science,
The Laboratory, Citadel Hill,
Plymouth PL1 2PB, UK
e-mail: pcre@wpo.nerc.ac.uk
*also at: Department of Biological Sciences,
University of Plymouth,
Drake Circus,
Plymouth PL4 8AA, UK

1. Houghton, J. T. et al. (eds) *Climate Change 1995* (Cambridge Univ. Press, 1995).
2. McCartney, M. S., Curry, R. G. & Bezdek, H. F. *Oceanus* **39** (2), 19–32 (1996).
3. Gamble, J. C. in *Long-term Experiments in Agricultural and Ecological Sciences* (eds Leigh, R. A. & Johnston, A. E.) 365–386 (CAB International, Wallingford, 1994).
4. Sir Alister Hardy Foundation for Ocean Science *Annual Report 1996* (Plymouth, 1997).
5. Joint, I. & Pomroy, A. *Mar. Ecol. Prog. Ser.* **99**, 169–182 (1993).
6. Myneni, R. B., Keeling, C. D., Tucker, C. J., Asrar, G. & Nemani, R. R. *Nature* **386**, 698–702 (1997).
7. Dickson, R., Lazier, J., Meincke, J., Rhines, P. & Swift, J. *Prog. Oceanogr.* **38**, 241–295 (1996).
8. Björge, E., Johannessen, O. M. & Miles, M. *Geophys. Res. Lett.* **24**, 413–416 (1997).
9. Hecht, J. *New Sci.* **156** (2103), 4 (1997).

Language steamrollers?

Jared Diamond's tabulation¹ of the number and degree of dissimilarity between the world's extant languages shows how unevenly this kind of human variation is distributed geographically. This is perhaps the greatest mystery of historical linguistics. If early Holocene linguistic upheavals are crucial for understanding human population genetics, it is no less crucial that we should be fully aware of the uncertain nature of the prehistoric facts that linguistic data can be used to reconstruct.

Diamond reports Bellwood and Renfrew's argument that major demographic upheavals ('steamrollers') at different times in the past 10,000 years, resulting from domestication of certain species of plants and animals, have erased "the products of previous tens of thousands of years of language evolution" in some parts of the world (for example, Europe) more strongly than in others (for example, New Guinea and 'native' California). But there is no generally accepted theory of language diversification^{2,3}; indeed, Renfrew and Bellwood's views are strongly contested⁴⁻⁶. Measuring the degree of isolation between human groups and estimating the length of time since their presumed separation (isolation) does not adequately predict either number of languages per unit area or their taxonomic (historical) diversity⁷.

These are a few of the reasons why we are reluctant to agree that modern New Guinea or aboriginal California can be taken as models for other places and times in history. Nobody knows what the linguistic diversity of Europe at the end of the Pleistocene was like and one can only guess at the linguistic situation in Europe even as late as the early Bronze Age, in the second millennium BC. Were one to accept Diamond's metaphor of the language steamroller as a heuristic device and his argument that speakers of (an unknown number of) early Indo-European languages moved "as a steamroller" over Europe, there is no way to tell whether these hypothesized migrants (or just the Indo-European language) had to flatten a sharply dissected linguistic landscape or rolled over a basically level field.

The solution to the mystery of historical linguistics will come only with better historical data and better sociolinguistic models⁸.

John Edward Terrell

New Guinea Research Program, The Field Museum,
Chicago, Illinois 60605, USA
e-mail: terrell@fmppr.fjnh.org

John Hines

School of History and Archaeology,
University of Wales, Cardiff CF1 3XU, UK

Terry L. Hunt

Department of Anthropology,

University of Hawaii at Manoa, Honolulu,
Hawaii 96822, USA

Chapurukha Kusimba

Department of Anthropology, The Field Museum,
Chicago, Illinois 60605, USA

Carl Lipo

Department of Anthropology,
University of Washington, Seattle,
Washington 98195, USA

The spread of Indo-European languages is often exclusively attributed to agricultural progress, whereas an alternative hypothesis, the domestication of the horse for warfare, tends to be ignored. Only a few years ago a good case was made for the steamrolling of undefended Old European cultures⁹ and languages by the first 'armoured vehicle', the horse^{10,11}, whose breaking to the bit coincided with proto-Indo-European languages geographically (Ukraine) and in time (at least 4,000 BC).

In anthropology we are witnessing the resurrection of hypotheses depicting our ancestors as fierce big-game killers, even cannibals¹², at the expense of more politically correct views of noble vegetarians or, at worst, scavengers. Could the spread of language be another aspect to this argument? Is the idea of peaceful, egalitarian cultures subjugated by male barbarians less palatable than their displacement by a superior agricultural technology?

Why were California and New Guinea spared a linguistic takeover? Could it be because the horse never got there? Perhaps the 'aggression' and 'technology' mechanisms for competitive exclusion of cultures are not incompatible, but rather occurred in different mixes, with the sword leading in Indo-Europe and the plough in Africa.

Wolfgang E. Sterrer

Bermuda Natural History Museum, PO Box FL
145, Flatts FLBX, Bermuda
e-mail: wsterrer@bbsr.edu

Diamond replies — By 'language steamroller' is meant the massive replacement of a region's languages by languages from the outside. Some military, technological, political or demographic advantage enables outsiders to impose their language or even to conquer or replace the original inhabitants. Terrell *et al.* question whether the suggested 'steamrollers' could be real, whereas Sterrer discusses their interpretation.

The evidence for steamrollers is overwhelming and familiar. In modern times, European overseas expansion resulted in English, Spanish, French and Portuguese replacing thousands of native languages of other continents. (Why else would this issue of *Nature* be printed for North American readers in English, rather than in a Native American language?) Abundant linguistic evidence testifies to many equally massive earlier replacements, of which the Bantu

expansion over sub-Saharan Africa and the Austronesian expansion over the tropical Pacific are particularly well attested by independent archaeological and genetic evidence.

In western Europe, all of the many modern languages except Basque belong to the Indo-European language family, which diversified during the past 5,000–10,000 years from an ancestral language spoken somewhere in western Asia. Do Terrell *et al.* believe all Europeans spoke a single tongue, Basque, until a few thousand years ago? In fact, preserved Etruscan and Iberian writing from Roman times, and residues of non-Indo-European words swept up into existing European languages, testify to the existence of many other non-Indo-European languages supplanted by Indo-European. Whether former language diversity in Europe was higher or lower than in modern New Guinea and aboriginal California is beside the point; the supplanting of that diversity still cries out for explanation.

As for the mechanism of that supplanting, Sterrer is correct: there are various types of advantage that can enable one group of people to conquer (or to impose their language on) another. Most known steamrollers can ultimately be related in some way to food production, because it is the agent that has produced the biggest effects on human population numbers and human societies in the past 10,000 years.

There is a long-standing and unresolved debate over whether the Indo-European steamroller was driven by the demographic advantages that West Asian food producers gained 9,000 years ago over Europe's original hunter/gatherer population, or whether the domestication of the horse (the jeep and Sherman tank of ancient warfare) became crucial after 6,000 years ago. As Sterrer notes, Gimbutas¹³ amassed strong arguments for the latter view, but Renfrew¹⁴ has also made a strong case for the former view. This is just one example of the many fascinating problems posed by the language steamrollers that jump out at us from language maps of the world.

Jared Diamond

Physiology Department, UCLA Medical School,
Los Angeles, California 90095-1751, USA

1. Diamond, J. M. *Nature* **389**, 544–546 (1997).
2. Hill, J. H. *Languages on the Land: Toward an Anthropological Dialectology* (Dept. of Anthropology, Indiana Univ., 1996).
3. Nettle, D. J. *Anthrop. Archaeol.* **15**, 238–403 (1996).
4. Hines, J. *Archaeol. Pol.* **34**, 183–197 (1997).
5. Zvelebil, M. J. *Eur. Archaeol.* **3**, 33–70 (1995).
6. Terrell, J. & Welsh, R. L. *Antiquity* **71**, 548–572 (1997).
7. Swadesh, M. in *The Origin and Diversification of Language* (ed. Sherzer, J.) 8–42 (Aldine-Atherton, Chicago, 1971).
8. Terrell, J. *et al.* *Curr. Anthropol.* **38**, 155–195 (1997).
9. Gimbutas, M. *The Civilization of the Goddess* (Harper Collins, San Francisco, 1991).
10. Anthony, D. *et al.* *Sci. Am.* **295**(6), 44–48A (1991).
11. Diamond, J. M. *Nature* **350**, 275–276 (1991).
12. Gibbons, A. *Science* **277**, 635–637 (1997).
13. Gimbutas, M. *The Balts* (Praeger, New York, 1963).
14. Renfrew, C. *Archaeology and Language: the Puzzle of Indo-European Origins* (Cape, London, 1987).

Paraflagellar rod is vital for trypanosome motility

African trypanosomes are protozoan parasites that cause sleeping sickness in man. In addition to the axoneme, their flagellum contains a large structure called the paraflagellar rod (PFR) whose function is unknown. We used an antisense RNA approach to produce a specific molecular ablation of the PFR structure. The mutant cells are paralysed, demonstrating that the PFR has an essential role in cell motility.

A PFR is found exclusively in the flagellum of three evolutionary ancient groups of protists: kinetoplastids¹, euglenoids² and dinoflagellates³. Its function has been the subject of much conjecture^{1,3} yet remains entirely cryptic. The major PFR components are two closely related proteins, PFR-A and PFR-C, sharing 60% amino acid sequence identity^{1,4-6}.

We cloned the near-full-length sequence of the *PFR-A* gene (nucleotides 78–1,758)⁷ in the reverse orientation into the strong trypanosome expression vector pHD451 (provided by S. Biebinger and C. Clayton, Heidelberg)⁸ and integrated into trypanosome DNA by homologous recombination⁹⁻¹¹. Transformants where the antisense construct was targeted to the inverted rRNA spacer or to the procyclin promoter did not produce a particular phenotype and the decrease in the amount of PFR-A protein was marginal. However, in one transformant, cells grew normally but sedimented to the bottom of the well and seemed paralysed. This mutant was cloned twice by limiting dilution and named *snl-1*.

Southern blot analysis of *snl-1* revealed that the plasmid was integrated within one of the two *PFR-A* loci (trypanosomes are diploid) owing to a rare homologous recombination event (not shown). Northern blot analysis (Fig. 1a) showed that only a tiny amount of *PFR-A* mRNA was present

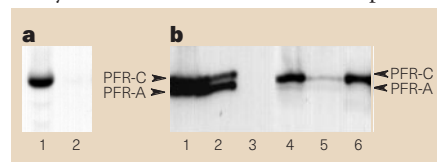


Figure 1 PFR-A expression is switched off in the *snl-1* cell line. **a**, Northern blot analysis of extracted total RNA from wild-type (lane 1) and *snl-1* (lane 2) trypanosomes hybridized with a *PFR-A* antisense RNA probe labelled with digoxigenin. Equal loading was confirmed by staining with ethidium bromide. **b**, Immunoblot analysis of wild-type (lanes 1–3) and *snl-1* (lanes 4–6) trypanosomes. Protein samples were probed with the L13D6 antibody. Lanes 1 and 4, total cell extract; lanes 2 and 5, pellet (detergent insoluble fraction); lanes 3 and 6, supernatant (detergent-soluble fraction).

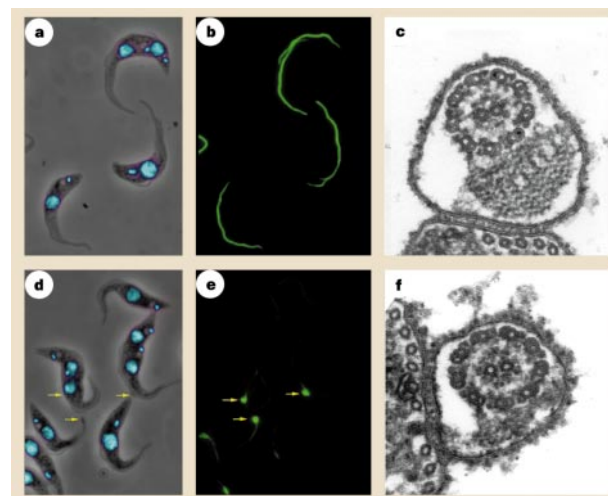


Figure 2 Morphological analysis of wild-type (**a–c**) and *snl-1* cells (**d–f**). Cells were stained with 4,6-diamidino-2-phenylindole, a DNA-intercalating dye staining the nucleus and the kinetoplast (shown in blue), superimposed on the phase-contrast image (**a,d**). **b,e**, Immunofluorescence pattern with the L13D6 antibody. Arrows show the dilation of the distal tip of the flagellum. **c,f**, Electron microscopic cross-sections of the flagellum of wild-type (**c**) and *snl-1* trypanosomes (**f**).

in the *snl-1* compared with wild-type cells. Western blots (Fig. 1b) of total cell lysates probed with L13D6, a monoclonal antibody recognizing both PFR-A and PFR-C (gift from Linda Kohl, University of Manchester) showed two bands of roughly equal intensities (PFR-A and PFR-C) in the wild-type cells (lane 1) but only PFR-C in the mutant trypanosomes (lane 4). This implies that *PFR-C* gene expression was not modified by the presence of antisense construct against *PFR-A*.

We extracted cells with detergent¹² to separate out the assembled cytoskeleton including the flagellum (Fig. 1b). In wild-type trypanosomes, the PFR proteins were found exclusively in the cytoskeleton (lane 2) and no soluble pool was detected (lane 3). In the *snl-1* mutant, the PFR-C protein was shifted to the soluble fraction (lane 6) and only a small amount was left in the cytoskeleton (lane 5). Therefore, in the absence of PFR-A, the PFR-C protein did not assemble in a cytoskeletal structure.

In immunofluorescence¹³, anti-PFR antibodies produced the expected bright signal in the flagellum of wild-type trypanosomes (Fig. 2a, b). This was markedly decreased in the *snl-1* mutant (Fig. 2d, e). Under phase-contrast microscopy, many *snl-1* cells exhibited a pronounced dilation of the distal tip of the flagellum (Fig. 2d, arrows) which stained strongly with the L13D6 antibody, indicating that non-assembled PFR-C entered the flagellum compartment and accumulated at the distal tip. Under the electron microscope we observed a major and specific ablation of the PFR in the *snl-1* mutant (Fig. 2c, f). Cross-sections of the wild-type flagellum revealed the typical PFR structure (Fig. 2c), composed of three regions and specific attachment points to the axoneme¹. In the *snl-1* flagella, the intermediate and distal regions of the PFR were missing and only a fraction of the smaller proximal region remained (Fig. 2f).

Wild-type trypanosomes swim actively *in vivo* and *in vitro*, the flagellum beating

from tip to base¹⁴. In the *snl-1* mutants, although the polarity was not affected, the frequency and the amplitude of flagellar beating were drastically decreased. Motility was so compromised that the *snl-1* cells sedimented rather than staying in suspension.

The paralysis of the *snl-1* mutant reveals an essential role for the PFR in flagellum and cell motility in trypanosomes. Our approach of integrating an antisense construct into one homologous gene locus suggests that position within the genome might be critical for the success of this method in trypanosomes. This could also be indicative of antisense interference with early processing. Production of such novel cellular phenotypes will find wide application in studies of parasites because motility functions have been implicated in events such as parasite–host attachment, surface-antigen trafficking and endocytosis.

**Philippe Bastin, Trevor Sherwin
Keith Gull**

School of Biological Sciences,
University of Manchester,
2.205 Stopford Building,
Oxford Road, Manchester M13 9PT, UK
e-mail: MBFPRBN@mh1.mcc.ac.uk

1. Bastin, P., Matthews, K. R. & Gull, K. *Parasitol. Today* **12**, 302–307 (1996).
2. Hyams, J. J. *Cell Sci.* **55**, 199–210 (1982).
3. Cachon, J., Cachon, M., Cosson, J. & Cosson, M. P. *Biol. Cell* **63**, 169–181 (1988).
4. Schlaeppli, K., Deflorin, J. & Seebeck, T. *J. Cell Biol.* **109**, 1695–1709 (1989).
5. Deflorin, J., Rudolf, M. & Seebeck, T. *J. Biol. Chem.* **269**, 28745–28751 (1994).
6. Moore, L., Santrich, C. & LeBowitz, J. H. *Mol. Biochem. Parasitol.* **80**, 125–135 (1996).
7. Bastin, P., Bagherzadeh, A., Matthews, K. R. & Gull, K. *Mol. Biochem. Parasitol.* **77**, 235–239 (1996).
8. Biebinger, S., Wirtz, L. E., Lorentz, P. & Clayton, C. E. *Mol. Biochem. Parasitol.* **85**, 99–112 (1997).
9. Ten Asbroek, A., Ouellette, M. & Borst, P. *Nature* **348**, 174–175 (1990).
10. Eid, J. & Sollner-Webb, B. *Proc. Natl Acad. Sci. USA* **88**, 2118–2121 (1991).
11. Blundell, P., Rudenko, G. & Borst, P. *Mol. Biochem. Parasitol.* **76**, 215–229 (1996).
12. Sherwin, T. & Gull, K. *Phil. Trans. R. Soc. Lond. B* **323**, 575–588 (1989).
13. Sherwin, T., Schneider, A., Sasse, R., Seebeck, T. & Gull, K. *J. Cell Biol.* **104**, 439–446 (1987).
14. Walker, P. J. *Nature* **189**, 1017–1018 (1961).

The story of ewe, again

Clone: The Road to Dolly and the Path Ahead

by Gina Kolata

Morrow/Allen Lane: 1997. Pp. 218. \$23, £15.99

Robert Bud

Next year we shall celebrate the sesquicentenary of the publication of one of the great Victorian paens to applied science, James Caird's pamphlet *High Farming under Liberal Covenants: The Best Substitute for Protection*. As Britain experimented with free trade for the first time, this polemic from the mid-nineteenth century advocated that, rather than relying on import duties, farmers should be encouraged to turn to science to enhance productivity and profits, promoting the new-technology agriculture known as 'high farming'.

How appropriate, then, that Dolly, 1997's scientific achievement of the year, was born at the Roslin Institute in the Scottish lowlands, home of Caird and the pioneers of agricultural improvement associated with the Highland and Agricultural Society.

In the eighteenth and nineteenth centuries, this rolling hill country was one of the most important locations for the conceptualization of the relationship between science, technology and humanity. The universities of Edinburgh and Glasgow nurtured some of the world's leading philosophers and scientists. Practice as exemplified by Caird and his peers was also among the most 'advanced' anywhere in the world.

Sheep and their products had been staples of human life in Scotland since the Bronze Age, and wool, milk, skins, hooves and meat were all exploited and refined. The mid-nineteenth century saw the emergence of tweed on the worldwide market, by which time the Clearances had brought sheep to the Highlands in place of man. The classic 800-page volume on the history of husbandry entitled *Sheep and Man* bears testimony to the intimate interrelationship of the two species over thousands of years and the roles of science and technology in mediation.

The Roslin Institute is descended from the department of research in animal breeding at the University of Edinburgh. The department was founded just a generation after the death of Caird.

By the time of Dolly's conception, the institute had already spawned a biotechnology company, PPL Therapeutics, which had



developed ways of expressing human genes in sheep. Flocks are milked in a barn a few miles from Edinburgh, and from their milk are extracted proteins such as the cystic fibrosis drug α -1-antitrypsin.

At present, of founder animals born by means of eggs injected with human genes which are then implanted into surrogate mothers, only 2–3% are transgenic, and it is possible only to add genes. Producing founder animals by nuclear transfer will be more efficient and will allow more sophisticated genetic modifications such as the deletion or replacement of genes.

As a practical man, Caird would have been interested in this demonstration of "science with practice", the motto of England's Royal Agricultural Society. The press release of 24 February 1997 from the Roslin Institute used the word 'technology' four times and 'scientific' or 'scientifically' only twice. As a Galloway farmer, Caird would have been pleased by the means of demonstration that the surrogate mother was not the genetic mother: the surrogate was a blackface ewe, whereas Dolly is clearly a white-faced Finn Dorset.

Gina Kolata is the science editor of *The New York Times* who covered the embargoed *Nature* paper describing what was the first successful cloning of an adult mammal (see *Nature* 385, 810; 1997). Her book might well have been a breathless recitation of recent exciting events, but has turned out to be far richer than such a parody would suggest.

Kolata was clearly taken aback by the pragmatic Scots tradition. "The circuit", she says, "that attracted the molecular biology and molecular genetics leaders did not include people who studied large farm animals like sheep, cows and pigs." She describes Roslin as being "outside the mainstream", and "not at an established university with an

aggressive public relations staff but at a tiny rural research institute where a guard dog named Buster roamed the grounds at night". The project leader, Ian Wilmut, was not in her newspaper's files, as they had never mentioned him before. What from one perspective was the heartland of science seemed, from another, to be quaintly removed from modernity.

The context of the achievement was surprising to those for whom cloning had already been defined in quite a different way from the technological challenge addressed outside Edinburgh. Throughout the twentieth century, and most particularly during the past 30 years, and above all in New York state, home of the Hastings Institute, cloning had come to be seen as a bioethical rather than a technological problem. Kolata subtly explores the way in which this definition came to pass.

Despite the speed of publication, Kolata has placed the development of Dolly within an original historical context. Not for her perfunctory gestures to the saga-cycle of genetics and biotechnology: Gregor Mendel and the story of his peas, soon followed by Francis Crick, James Watson and the story of the double helix. Nor does she resort to appealing to the Frankenstein archetype. Rather, she has addressed the complex development of embryology through the scientific styles of three generations of scientists since the beginning of the twentieth century, and their varying engagements with philosophy. Benefiting from Victor Hamburger's work on Hans Spemann, Kolata highlights the philosophical concerns of Spemann and his teacher Boveri early in the century. In his *Embryonic Development and Induction*, Spemann suggested that the nucleus from a mature cell could be transplanted into an egg.

Kolata then points to the growth of



rewards for high-prestige science after the Second World War which led to a research style, particularly in the United States, very different from that of pre-war Europe. She contrasts the self-consciously philosophical style of the Germans with the peer-review-panel-conscious, grant-dependent Americans Robert Briggs and Tom King, who in 1951 managed to clone frogs' eggs from embryos.

Kolata provides a nice account of the specialization that led to a distinctive bioethics movement in the United States with its roots in concerns about the rationing of health care (kidney dialysis in particular), medical experiments (the Tuskegee syphilis experiment and hepatitis B) and the new genetic medicine. She shows how, in the deeply distrustful environment of the United States in the late 1960s, these diverse issues were brought together, and provides a valuable stimulus to further thought about the period.

To many at the time it seemed that science would go on remorselessly, despite the fledgling concern for bioethics. In 1978 the science writer David Rorvik claimed a scoop: the report of the first cloning of a human being. Coming shortly after reports of recombinant-DNA technology, it seemed to the public all too likely. Films such as Ira Levin's *The Boys from Brazil* explored the implications of cloning. The discussion of what ultimately proved a fantasy crystallized the opinion that this was a premature panic.

Yet around the same time the distinguished embryologist Karl Illmensee was becoming an academic celebrity for apparently succeeding in transferring the nucleus of one mouse into the embryo of another. He then became the centre of controversy when it was claimed that he had misrepresented his experiments. Kolata has clearly spoken to many of the participants in the Illmensee affair. Was Illmensee's work genuine? She explores its history in great detail, not resolving the question but pointing out the way it came to raise ethical issues in the context of cloning.

Kolata then considers the case of Steen Willadsen, who worked with farm animals in the Agricultural Research Council's Unit on Reproductive Physiology and Biochemistry at Cambridge. He managed to create chimerae by successfully transferring nuclei not just between animals but also between species. Unlike his US counterparts, he was not under pressure to publish. When he did, he chose *Veterinary Record* as well as *Nature*. He wanted to engage in "great absorbing endeavours".

Kolata was so involved personally in the third part of her account that it seems almost autobiographical at times. It describes how the technologically inclined Keith Campbell, working with Ian Wilmut, who had preceded Willadsen at Cambridge, produced Dolly

and then marketed her. The science, which is clearly described, is a matter of record, so the distinctive aspect of the book is perhaps the story of the interaction between the pragmatic Scottish context and the US bioethics community.

Kolata describes the media blitz that hit the Roslin Institute. The institute had been preparing for three months to cope with considerable public response and, on the Sunday when the British newspaper the *Observer* broke the story, six people were in place to handle enquiries. But the scale of interest was far greater than anything anticipated. The ethicist Ronald Munson is quoted as exclaiming: "Here we have this incredible technical accomplishment and what motivated it? The desire for more sheep milk of a certain type."

Kolata addresses the implications for philosophy and medicine and cites both enthusiasts and sceptics. But potential practical uses of the technique are not her concern; rather, her focus, like that of many critics, is the possibility of human clones. Here I think she is less successful at grappling with the issues, which are admittedly still ill-formed.

Is *Clone* a magisterial reflection on how Dolly has brought to fruition the century-old dream of embryology? Clearly not: Gunther Stent's name acquires a 'd', Driesch

has become Dreisch, characters have been omitted, and so on. Yet the extraordinary worldwide interest in cloning has stimulated a reflective and timely essay. In the heat of the moment, Kolata has given us a mature reflection on the shifting borders between technology, philosophy, public concern and science; and she has managed to tell some good stories, too. □

Robert Bud is at the Science Museum, Exhibition Road, London SW7 2DD, UK.

Homage to Earth

Gaia's Body: Toward a Physiology of Earth

by Tyler Volk

Copernicus: 1998. Pp. 269. \$27, £19

Peter Westbroek

It is almost two decades since the inventor James Lovelock brought an ancient goddess back to life. She was Gaia, mother Earth, who, in Lovelock's view, had made our planet a comfortable home for life for the past four billion years.

The 'Gaia hypothesis' has always aroused controversy among the scientific community. The idea of a 'homeostatic superorganism' was accused of being poorly defined and

The shell of Aphrodite

Since antiquity, molluscs — in particular the nautilus — have been associated with Aphrodite, the goddess of love, as a symbol of the female pudenda, sexual desire and aphrodisiacs. Although the female paper nautilus forms a shell, the male, as shown in this nineteenth-century lithograph, does not; to fertilize the female, one of the tentacles of the male separates from its body and swims towards her as a giant penis. The mollusc is one of several animals that have sneaked into Christian Ratsch's *Plants of Love* (Ten Speed Press, \$19.95, pbk), a leisurely, visually enticing romp through the history, botany, chemistry, pharmacology and identification of various aphrodisiac plants. User instructions are also included. *Caveat emptor!*



impossible to falsify. Furthermore, because Gaia is unique and hence not subject to natural selection, some evolutionary biologists could not perceive an organizing principle or mechanism by which global homeostasis could be explained. And worst of all, Gaia received an extra mystical flavour as it was adopted by the New Age movement.

Over the years, Lovelock and his collaborators have done their utmost to clarify the concept and make it suitable for scientific research. Despite its flowery title, *Gaia's Body* is a further attempt at demystification, and, I dare say, is a major step forward.

For years Tyler Volk has been studying the carbon cycle, crop growth for the US space agency NASA's closed systems, and the climatic effects of evolutionary change. It is this personal mix of system dynamics and Earth and life science that makes him talk about 'Gaia' rather than 'biosphere'.

This book has convinced me that Volk is right: it is not Gaia but biosphere that is the woolly concept. Biosphere conveys little more than a vague notion of where life can be found. Gaia, as described by Volk, relates to a dynamical system with a particular behaviour and clearly defined boundaries. It comprises not only life itself, but also the atmosphere, soils and ocean with which it interacts. This definition identifies the system well, as material is circulating far more rapidly within it than is exchanged with its environment, the deeper Earth and outer space. These exterior domains set the boundary conditions for Gaia: they provide energy and topographical relief, supply basic nutrients, and act as the ultimate recipient for refuse.

Volk argues convincingly that Gaia may be compared to a physiological system, even if it does not evolve in the strict Darwinian sense. Atmospheric and ocean circulation and the continent-draining water cycle intimately connect all its parts. For example, Keeling's annually oscillating curve of atmospheric carbon dioxide in Hawaii reflects the alternating photosynthetic and respiratory cycles of continental vegetation in the Northern Hemisphere. In addition, the finer details of this curve record the multiple interactions of all the additional components of the carbon cycle.

Another case in point is the recycling of essential nutrients, which is high for oceanic phosphate. This property of Gaia sets the stage for a huge amplification of biological activity on Earth. Also fascinating is Volk's account of how the cycles of nitrogen and phosphate are coupled in the ocean, in both sea water and organisms.

These examples, and many others discussed by Volk, indicate that Gaia is organized, at least to some degree, despite the evolutionary biologists' theoretical reservations. Volk argues that, instead of natural selection among other planetary systems, it

is the proliferation of life within the constraints of the Gaia system that automatically brings about the organization.

What are the functional parts of Gaia? Are they the ecosystems, the hierarchy of systematic taxa, or the cycles of elements? Volk emphasizes a subdivision in what he calls biochemical guilds: groupings of organisms that perform similar functions, such as photosynthesis, respiration, sulphate reduction or nitrogen fixation. He then describes some important species-transcending "molecules that run the world", such as chlorophyll, Rubisco and glutamine synthetase, and explains how they combine to form both the cellular and the Gaian metabolic web.

This leads him to the idea, developed by G. J. Williams in his book *The Molecular Biology of Gaia* (Columbia University Press, 1996), that the molecular controls on the activities of these universal enzymes are also responsible for fine-tuning the global biogeochemical cycles.

The attraction of Williams's hypothesis is obvious. It allows us to study Gaian metabolism directly from its molecular biological underpinnings and to ignore the ecological level of biological organization. It may even work for the cycling of some critical nutrients, such as phosphate and nitrogen.

But what about the production of the most important carbon sink on Earth, calcium carbonate, virtually all of which is made by organisms. In the open ocean more than 200 species of coccolithophore algae and foraminifera make calcium carbonate, and the amount produced depends on their success in proliferating. They compete with thousands of other organisms, are subject to intensive grazing as well as bacterial and viral infections, and have to cope with the vagaries of ocean circulation. However fuzzy and complex the ecosystem concept may be, it cannot be ignored.

Despite this and other minor shortcomings, *Gaia's Body* is an outstanding contribution to global ecology. Its main virtue is that it brings the Gaia concept to the heart of science. By defining clearly the boundaries of the Gaian system and placing biogeochemistry in that context, the concept loses much of the fuzziness it may have had for some scientists.

The book is very well written and easy for a broadly educated audience to follow. Although I am familiar with much of the content, the details are knitted together so well that I read it as an exciting novel. I am pleased to see that Lovelock's invention is getting such an excellent follow-up. It is to be hoped that *Gaia's Body* will further the acceptance of Gaia by the scientific community. □

Peter Westbroek is at the Leiden Institute of Chemistry, Gorlaeus Laboratories, University of Leiden, PO Box 9502, 2300 RA, Leiden, the Netherlands.

New in paperback

Scientific and Ordinary Action: Ethnomethodology and Social Studies of Science

by Michael Lynch

Cambridge University Press, £14.95, \$21.95

An empirical approach to the sociology of scientific knowledge that attempts to avoid the pitfalls of 'scientism' and 'foundationalism'.

Science, Nonsense, and Nonsense: Approaching Environmental Literacy

by Michael Zimmerman

Johns Hopkins University Press, \$14.95, £12.50

"Zimmerman's account of [the battle between creationists and Darwinian evolutionists] is fascinating and at times frightening", John Adams, *Nature* 381, 125 (1996).

Krakatau: The Destruction and Reassembly of an Island Ecosystem

by Ian Thornton

Harvard University Press, \$18.95, £12.50

"Exciting, information-rich... will provide hours of truly pleasurable reading and also clear and deep insights into the fascinating biology of islands, the struggle to construct comprehensive theories for difficult processes and the techniques for preserving species diversity", Lawrence B. Slobodkin, *Nature* 381, 205 (1996).

The Snow Leopard

by Peter Matthiessen

Vintage, £5.99

Matthiessen recounts his 1973 trip with the field biologist George Schaller to the Himalayas in search of the rare snow leopard, and tells of the inner spiritual journey that it engendered.

Higher Superstition: The Academic Left and Its Quarrels with Science

by Paul R. Gross and Norman Levitt

Johns Hopkins University Press, \$16.95, £14

Since its publication in 1994, this book has been at the centre of the 'science wars'. "Convincing, and therefore frightening.... This is a conversation that needs many participants — including, most especially, those who have chosen to ignore or shrug off what has become an embarrassment to serious scholarship", Donald Kennedy, *Nature* 368, 409 (1994).

Hunting Down the Universe: The Missing Mass, Primordial Black Holes and Other Dark Matters

by Michael Hawkins

Abacus, £7.99

"A good yarn. But I think the actual situation is more prosaic and less polarized.... In due course, as more data on quasar fluctuations accumulate, [the author's] hypothesis is very likely to be disproved to the satisfaction of any sapient individual", William H. Press, *Nature* 388, 138 (1997).

Probing the sea

Molecular Approaches to the Study of the Ocean

edited by Keith E. Cooksey
*Chapman and Hall: 1998. Pp. 549. £110,
 \$199.95*

N. G. Carr

About 20 years ago, some biological oceanographers, not least the late Ian Morris, began to argue that the application of molecular biology to the study of oceanic phytoplankton would be the next impetus to move the subject along, following the success of chemical and physical techniques such as radioactive isotope flow and fluorescence analysis. The recognition around that time of the microbial loop, in which a large proportion of photosynthetic productivity is recycled at the bacterial and protist level, would have encouraged such optimism.

Compared with clinical research, for example, oceanographic applications were rather slow in being optimized, but the use of molecular biology is at last changing our understanding of ocean microbiology in terms of what is there and what it does. *The Molecular Ecology of Aquatic Microbes* (Springer, 1995) was the first collection of articles on the subject, but *Molecular Approaches to the Study of the Ocean* also emphasizes subjects usually associated with aspects of marine biology such as toxins and adhesive proteins. It is a moot point whether the benefits of its broad coverage compensate for a degree of loss of focus.

The development of a molecular phylogeny based on 16S RNA gene sequence has revolutionized ideas of bacterial relatedness and allowed the first rational understanding of bacterial evolution. An excellent chapter by DeLong discusses the application of molecular phylogenetics to marine microorganisms including protists and some metazoans. A valuable part of this is a comparison of some of the methods of handling the sequence data that lead to the now familiar unrooted tree expression of relatedness. This approach has produced an analysis of the bacterial community structure within the open ocean. The myriad organisms seen microscopically and largely unidentified can now be related to the known microbial world. An important outcome of this work, which has relevance to microbiology as a whole, has been the discovery that the Archaea appear to be widespread in the ocean and not, as previously thought, confined to extreme environments.

One of the difficulties of working with, and writing about, marine bacteria and phytoplankton is relating their microbiological interest to their contribution to the understanding of the overall process of biological oceanography. In a long chapter with the

deceptively simple title of "Molecular approaches to microbial biomass estimation", Karl and Dobbs achieve this splendidly. They combine details of how the current experimental strategies have evolved with a feel for the sheer scale of the microbially driven biogeochemical cycles in what they describe as the largest contiguous habitat on Earth.

The feature of molecular biology that makes it so effective in unravelling biological oceanography is the specificity that derives from nucleic-acid sequence probes and immunological recognition of proteins. The article by Zehr and Hiorns exemplifies this and builds on successful applications for measuring aspects of bacterial nutrition and rates of phytoplankton photosynthesis to suggest a wide range of cellular activities open to molecular examination and of use in determining the metabolic status of different components of microbial populations.

The book has 27 chapters, so perhaps it is not surprising that several describe fields of study showing modest progress and that one or two are made up largely of previously published material. For the most part though, the book comprises relevant, well-written accounts that cover the molecular analysis and ecology of the oceans. Other advanced techniques, such as flow cytometry and developments in fluorescence analysis, are discussed in a way that indicates that the authors and editor appreciate that progress will usually involve the integration of disciplines.

Most researchers in biological oceanography will find things of value in this well-produced book, and some of the chapters provide for a wider readership an excellent introduction to this critical area of environmental biology. □

N. G. Carr is in the Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, UK.

Patchy coverage

Climate Process and Change

by Edward Bryant
*Cambridge University Press: 1997. Pp. 209.
 £45, \$59.95 (hbk); £16.95, \$22.95 (pbk)*

Thomas F. Stocker

The science of Earth's climate system has gone from being the preoccupation of a handful of pioneers to a large international research effort. Its unique complexity means that only the combination of high-quality research across disciplinary boundaries will generate the innovative results needed for progress. So a book intended for use in undergraduate courses is timely.

In *Climate Process and Change*, Edward Bryant seeks to present in simple terms the workings of the climate system and its

changes. He does so in three parts, on processes, change and impact, in just over 200 pages, too few for in-depth discussion. Readers would appreciate references to more specialized sources or excellent textbooks such as *Physics of Climate* by J. P. Peixoto and A. H. Oort (American Institute of Physics, 1992) or *Paleoclimatology* by T. J. Crowley and G. R. North (Oxford University Press, 1991).

More than half the book focuses on the past century, and the question of global warming is a recurring theme, presented here in a rather one-sided manner. Students reading this text would end up with the "acquired bias from previous instruction" that Bryant sets out to correct.

Central to our understanding of climate processes are the natural sciences, whose language is equations. This book contains very little — and often sloppy — physics, and the few equations are hardly useful; for example, $y = bx^n$ (where $n \neq 1$) is explained by "Many climatic processes behave like this and have a value of n between 1.0 and 2.0".

A chapter on the transport of heat and matter introduces the reader to large-scale meridional air flow — the concept of transient eddies is not mentioned — and to urban climate and "mobile polar highs". We learn that this concept, unknown to dynamical meteorologists, explains the atmospheric circulation for the climatic states of the glacial, today and in the future. A narrow and often incorrect view of ocean circulation and processes follows.

In the discussion of Pleistocene change, there is no mention of the rapidly accumulating results from various novel proxy indicators extracted from marine sediments, corals and tree rings. There is a fairly complete account of the causes of climate change but stating that it is impossible to attribute temperature change during the past century to different causes ignores recent promising advances in statistical analysis.

The book closes with two interesting chapters on the impact of climate change on health and ecosystems. I am happy to learn that climate-related diseases "may not be the biggest threat to human health" and that food contamination at a "fast-food outlet or wedding reception" may be more dangerous.

In his epilogue Bryant says that "if greenhouse warming is a reality, then its arrival may well be timely, because it has the potential to compensate for the strong cooling that may be a natural part of the ocean-atmosphere system, under the heavy atmospheric dust and sulphate aerosol loadings presently being stimulated by human activity". I prefer his alternative view, earlier in the book: "as with all human tampering with climate, there are bound to be negative ramifications". □

Thomas F. Stocker is in the Division of Climate and Environmental Physics, Physics Institute, University of Bern, 3012 Bern, Switzerland.

Three-dimensional preservation of algae and animal embryos in a Neoproterozoic phosphorite

Shuhai Xiao*, Yun Zhang† & Andrew H. Knoll*

* Botanical Museum, Harvard University, 26 Oxford Street, Cambridge, Massachusetts 02138, USA

† College of Life Sciences, Beijing University, Beijing 100871, People's Republic of China

Phosphorites of the late Neoproterozoic (570 ± 20 Myr BP) Doushantuo Formation, southern China, preserve an exceptional record of multicellular life from just before the Ediacaran radiation of macroscopic animals. Abundant thalli with cellular structures preserved in three-dimensional detail show that latest-Proterozoic algae already possessed many of the anatomical and reproductive features seen in the modern marine flora. Embryos preserved in early cleavage stages indicate that the divergence of lineages leading to bilaterians may have occurred well before their macroscopic traces or body fossils appear in the geological record. Discovery of these fossils shows that the early evolution of multicellular organisms is amenable to direct palaeontological inquiry.

Most of the fossils that document the first 85% of evolutionary history are microscopic. Not until the Phanerozoic eon (<544 Myr BP) do the remains of large animals, algae and, later, plants become conspicuous constituents of the sedimentary record. The most important biological event that connects these palaeobiologically distinct eras is the evolution of complex multicellularity in eukaryotes. Multicellular organisms arose at least six times: in animals, fungi and several groups of algae¹. Microscopic remains of uncertain systematic affinities occur in rocks as old as 1,800–

2,100 Myr BP² and cellularly preserved microfossils of red, green and stramenopile (brown and related) algae indicate that multicellularity was achieved in these groups by about 1,000 Myr³. Multicellularity may have evolved comparably early in minute ancestral animals, but until now any pre-Ediacaran animal history has been contentious and thought by many to be unrecognized and perhaps unrecognizable by palaeontologists.

Phosphorites of the Doushantuo Formation in southern China contain three-dimensionally preserved fossils that record in exquisite

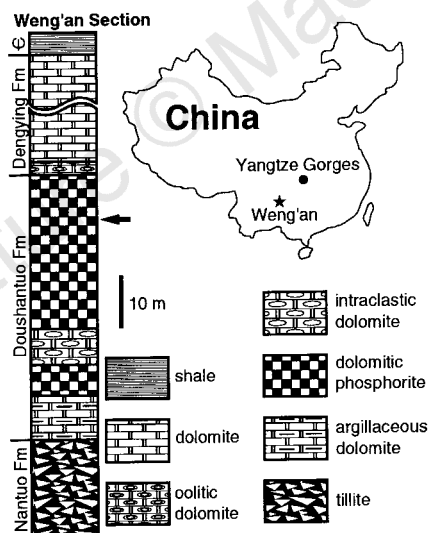
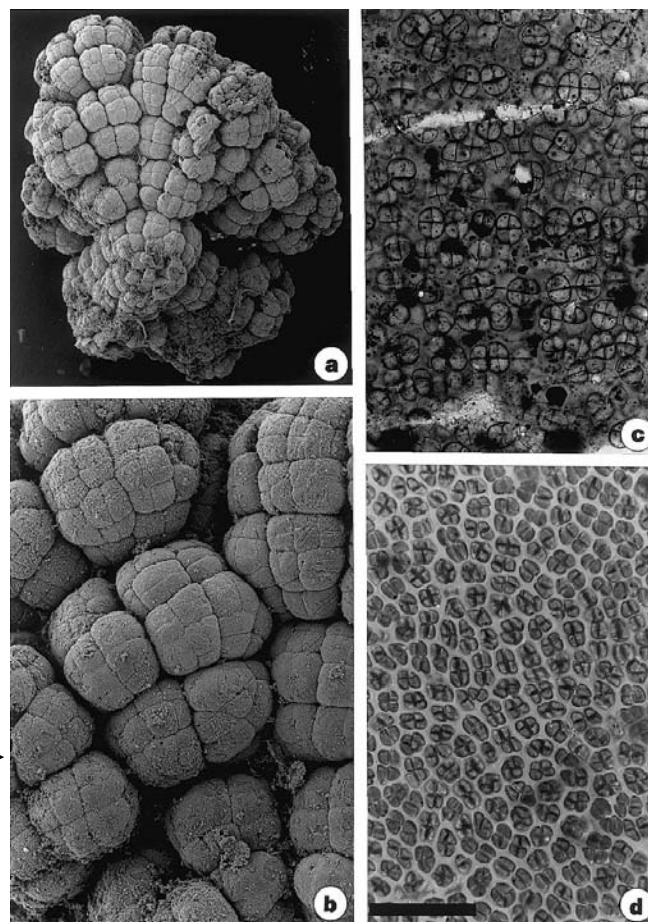


Figure 1 Location and generalized stratigraphy of the fossiliferous Weng'an section in Guizhou Province, South China. The arrow indicates the principal stratigraphic horizon containing phosphatized algae and embryos. Fm, Formation.

Figure 2 Algal thalli from the Doushantuo phosphorite and the modern bangiophyte red alga, *Porphyra suborbiculata*. **a–c**, Thalli from Doushantuo phosphorite. **a**, **b**, Scanning electron micrographs of a thallus composed of cuboidal cell packets similar to those of modern chlorosarcinacean green algae; **b**, the upper left quadrant of **a** at higher magnification. **c**, Photomicrograph of cruciate cell tetrads embedded in a foliose thallus; **d**, Carposporangia within the thallus of living *Porphyra*. Scale bar (in **d**): 200 μ m for **a**; 50 μ m for **b**; and 100 μ m for **c** and **d**.



cellular detail the anatomy and reproductive biology of diverse multicellular algae that lived in the late Neoproterozoic ocean. Doushantuo phosphorites also contain large populations of globular fossils which we interpret to be embryos of early animals. The quality of preservation and evolutionary importance of these fossils rival those of younger Lagerstätten such as the Burgess Shale^{4,5} or Rhynie Chert^{6,7}, and shed unprecedented palaeontological light on the early evolution of multicellular organisms.

Geological setting and fossil preservation

In its type area along the Yangtze Gorges, the Doushantuo Formation comprises a 250-m succession of carbonates, shales and phosphatic shales that lie disconformably above the glaciogenic rocks of the Nantuo Tillite and conformably beneath the carbonates of the Dengying Formation^{8,9} (Fig. 1). Dengying successions contain rare Ediacaran fossils^{8,10} and, in their uppermost part, basal Cambrian shelly fossils¹¹. Depositional age is only broadly constrained by U–Pb dates on volcanic rocks: tuffs in the underlying Liantuo Formation date from 748 ± 12 Myr⁸, whereas bentonites in the overlying Cambrian Zhongyicun Formation have an older age

limit of 539 ± 34 Myr¹². Diverse acritarchs and a distinctive carbon-isotopic signature, however, allow unambiguous correlation with better-dated successions elsewhere, indicating that Doushantuo sediments accumulated 570 ± 20 Myr ago¹³. Most diverse Ediacaran assemblages are younger than ~ 550 Myr^{14,15}, although frond-like fossils in Newfoundland occur in association with 565 ± 3 Myr ash beds¹⁶. Globally, acritarch assemblages of the type found in Doushantuo rocks occur in strata that underlie Ediacaran macrofossils. Thus available data indicate that the Doushantuo Formation is older than the diverse Ediacaran assemblages of South Australia and northern Russia, and may antedate all known Ediacaran-type assemblages, except for the simple centimetre-scale discs reported from pre-Varanger rocks in northwestern Canada¹⁷.

About 600 km to the southwest of the Yangtze Gorges, in the Weng'an region of Guizhou Province (Fig. 1), the Doushantuo Formation has similar acritarchs but markedly different lithologies. Here, Doushantuo sections are only 40–50 m thick and consist predominantly of phosphorites that lie unconformably above older metasediments and, locally, thin Nantuo diamictites¹³ (Fig. 1). Above a thin (0.5 m) basal conglomerate and up to 12 m of massive

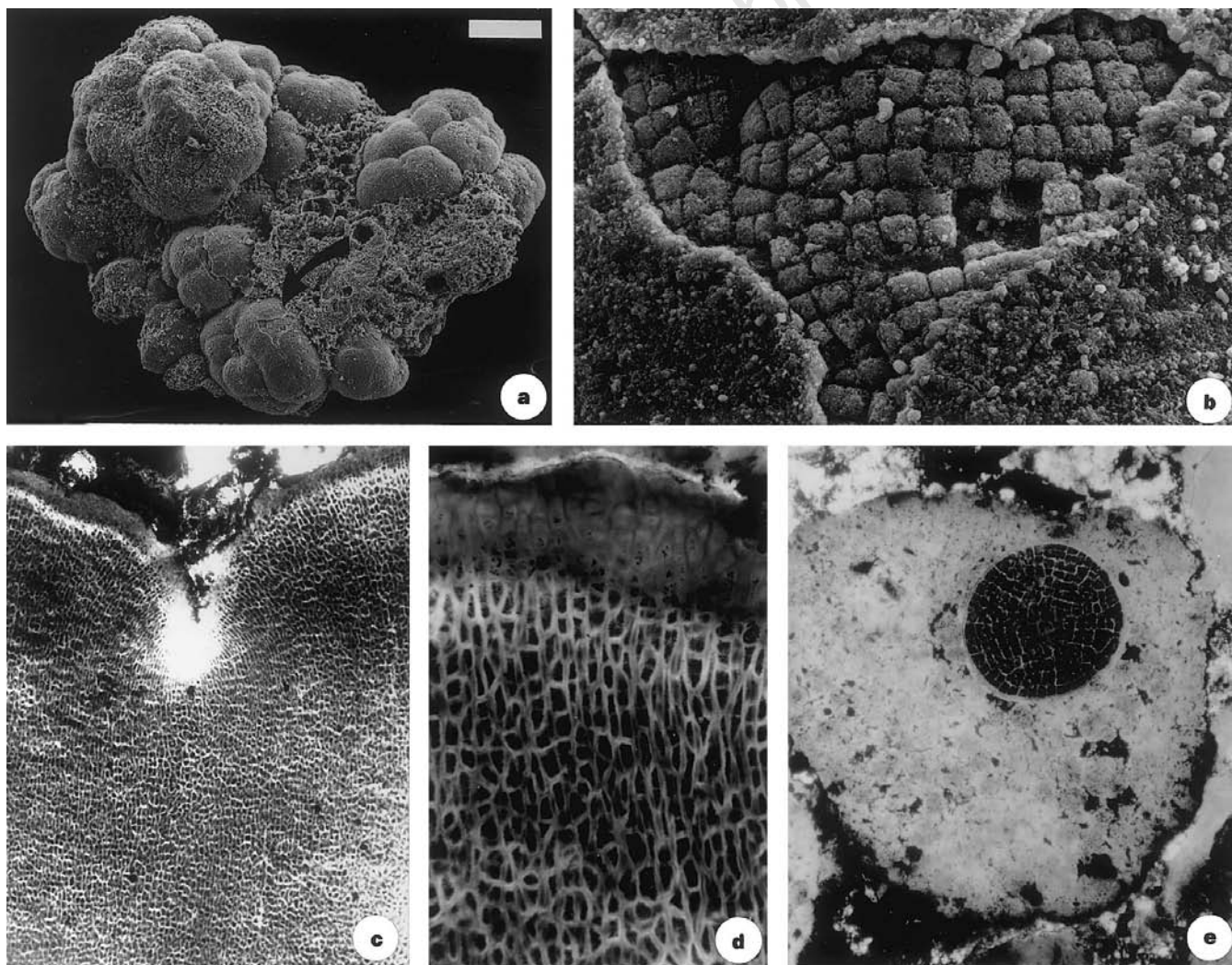


Figure 3 Parenchymatous and pseudoparenchymatous thalli from the Doushantuo phosphorite. **a**, Scanning electron micrograph showing the three-dimensional morphology of a pseudoparenchymatous thallus. **b**, Higher-magnification view of a portion of the same thallus (the arrow in **a** denotes position) showing the cell-surface pattern. **c**, Photomicrograph of a phosphatized thallus in thin section, showing pseudoparenchymatous 'cell fountain' anatomy. **d**, Higher magnifica-

tion of the same thallus, showing the details of the cell arrangement; the apparently distinct cells at the top of the figure reflect taphonomic and not anatomical differentiation. **e**, Parenchymatous thallus preserved within a phosphatic intraclast. Scale bar (in **a**): 140 μ m for **a**; 7.5 μ m for **b**; 55 μ m for **c**; 20 μ m for **d**; and 100 μ m for **e**.

dolostone lie 6–18 m of phosphatic mudstones and subordinate grainstones. The phosphorites exhibit parallel to slightly undulose lamination and are interbedded with fine-grained siliciclastic rocks in the lower part of the unit. Phosphatic beds coarsen upwards, become increasingly dolomitic, and locally contain stromatolites. Phosphatic and dolomitic grainstones, the latter characterized by cross-bedding and phosphorite nodules, are capped by a coarsely conglomeratic unit associated with subaerial erosion. Above this exposure surface, the cycle roughly repeats, with silicified phosphatic dolomitic and dolarenites of the uppermost Doushantuo Formation abruptly overlain by oolitic dolostones of the basal Dengying Formation. The environment of phosphorite deposition is interpreted as a shallow subtidal platform subject to episodic storms.

Fossils occur in both the lower and upper phosphorite sequences. *In situ* collophane crusts contain oriented algal thalli along with generally uncompressed microfossils and fine-grained organic detritus. Fossils also occur in interbedded, locally derived phosphatic grainstones and gravelstones. Phosphatization was penecontemporaneous with deposition, as shown by the local reworking of phosphatized fossils to form thin bioclastic sandstones¹³.

The chemistry of phosphate permineralization is not well understood, but it has been shown that soft tissues can become phosphatized within days of death¹⁸. In general, phosphogenesis requires both that phosphate be supplied to surface sediments in relatively high concentrations and that physical and/or biological processes operating within sediments further increase pore-water PO_4^{3-} concentration beyond the point of saturation with respect to ap-

tite^{19–21}. Divergent models have been advanced to explain the genesis of bedded phosphorites on ancient shelves and platforms, but most share several features. High biological productivity, commonly related to the upwelling of nutrient-rich deep water, is commonly invoked, as is an ‘iron-pumping’ mechanism in which particulate FeOOH scavenges PO_4^{3-} from the water column and delivers it to sediments, where it is reduced, liberating PO_4^{3-} and increasing pore-water PO_4^{3-} concentration. Dysaerobic bottom waters are thought to facilitate phosphogenesis because of their elevated PO_4^{3-} concentration. On ancient shelves and platforms, delivery of suboxic waters has commonly been attributed to transgression^{19–21}, and the two episodes of phosphorite deposition recorded in Doushantuo successions certainly reflect marked transgressions across the Yangtze Platform¹³. Continental weathering in warm, postglacial environments may also have contributed to regional phosphate enrichment. Low rates of sedimentation, the absence of bioturbation, and extensive cover by microbial mats and thalloid algae would all have facilitated phosphate precipitation.

Multicellular algae

Zhu *et al.*²² first reported fossils in Doushantuo phosphorites, but it was Zhang^{23,24} who established the superb preservation and diversity of this assemblage. In continuing collaborative investigations, we have documented eight cyanobacterial and 31 acritarch taxa in the formation as a whole, as well as eight formally named and numerous other fragmentary remains of multicellular algae¹³. Several other authors have reported evidence of animal remains in Doushantuo rocks. Rare triact spicules in Doushantuo cherts may record early

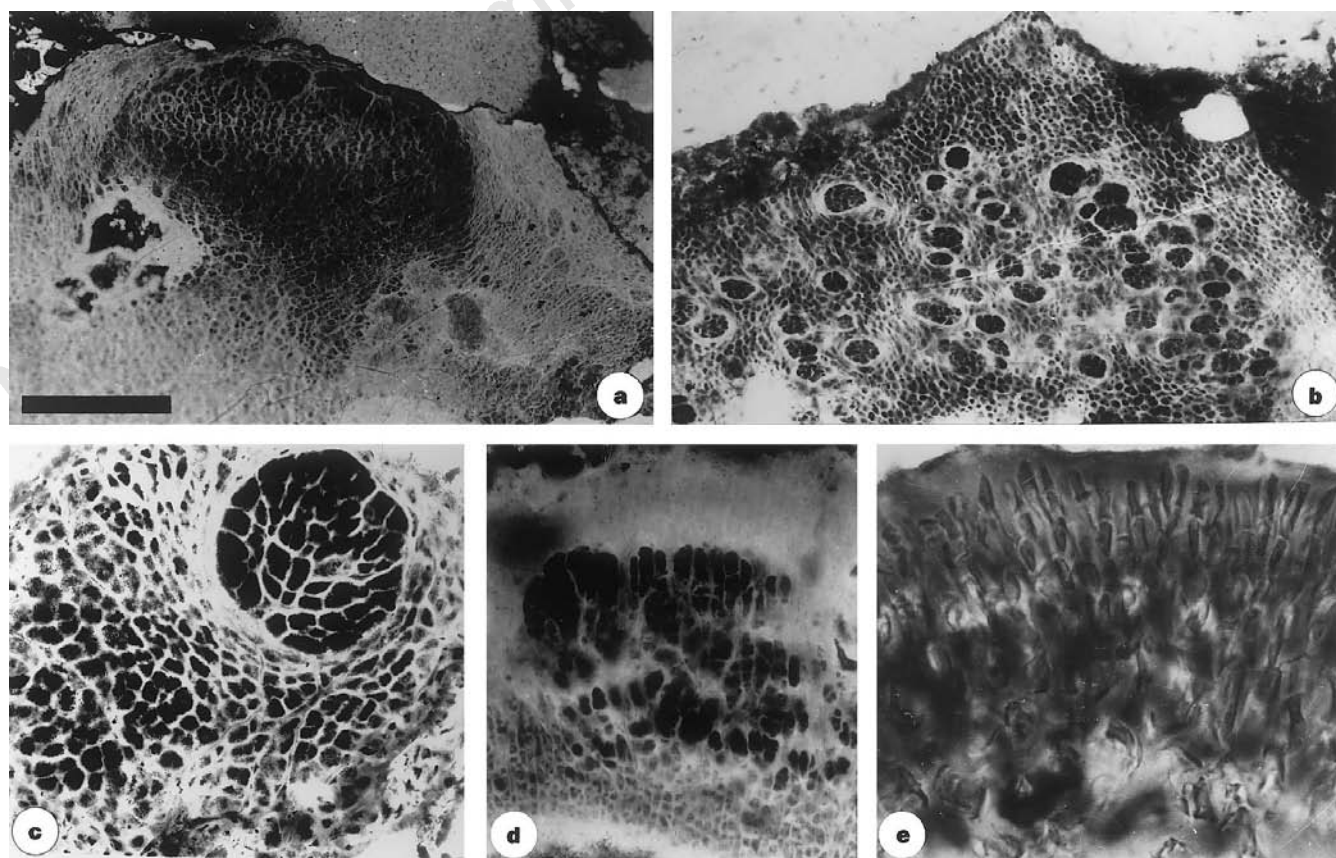


Figure 4 Reproductive structures in Doushantuo thalli and a modern red alga. **a–d**, Doushantuo thalli: **a**, Conceptacle in which carposporangia (clusters of dark, large cells) arise from gonimoblast filaments or supporting cells. **b**, Carposporangia (packets of dark cells) embedded in an anatomically preserved thallus. **c**, Higher-magnification view of carpospores within encompassing vegetative

tissue. **d**, Linearly arranged, dark, elongate cells interpreted as possible spermatangia. **e**, Spermatangial sori of the modern florideophyte red alga *Gracilaria* sp., for comparison with **d**. Scale bar (in **a**): 100 μm for **a**, **b** and **d**; 50 μm for **c**; and 30 μm for **e**.

sponges^{8,9}, but structures originally interpreted as microburrows²⁵ appear to be oblique sections through large, multilamellate cyanobacterial filaments¹³.

Doushantuo algal thalli range from simple colonies of undifferentiated cells to erect, branching forms characterized by tissue differentiation and specialized reproductive structures. Among the simplest multicellular entities are stacked cuboidal cell packets (Fig. 2a, b) of a type referred to as “incipient tissues”²⁶. Very similar structures occur today in several clades of green algae^{27,28}. Discoidal parenchymatous thalli (Fig. 3e) also occur in Doushantuo phosphorites. These fossils are morphologically simple as well, but in this case deceptively so; parenchymatous growth represents an evolu-

tionary departure from the plesiomorphic states in red, green and brown algae, where developmentally regular multicellularity is based fundamentally on filamentous growth^{27,28}.

Pseudoparenchymatous thalli, formed by the coordinated growth of closely packed filaments, are abundant in the Doushantuo assemblage (Fig. 3a–d). A distinctive characteristic of pseudoparenchymatous growth is the cell fountain, comprising vertical rows of cells that have expanded and diverged upwards to form a fountain-like array in longitudinal section. Cell fountains are common features of Doushantuo thalli, and similar features typify several clades of living red algae, although they also occur in brown algae such as *Ralfsia*²⁶. Rhodophytic affinities are sup-

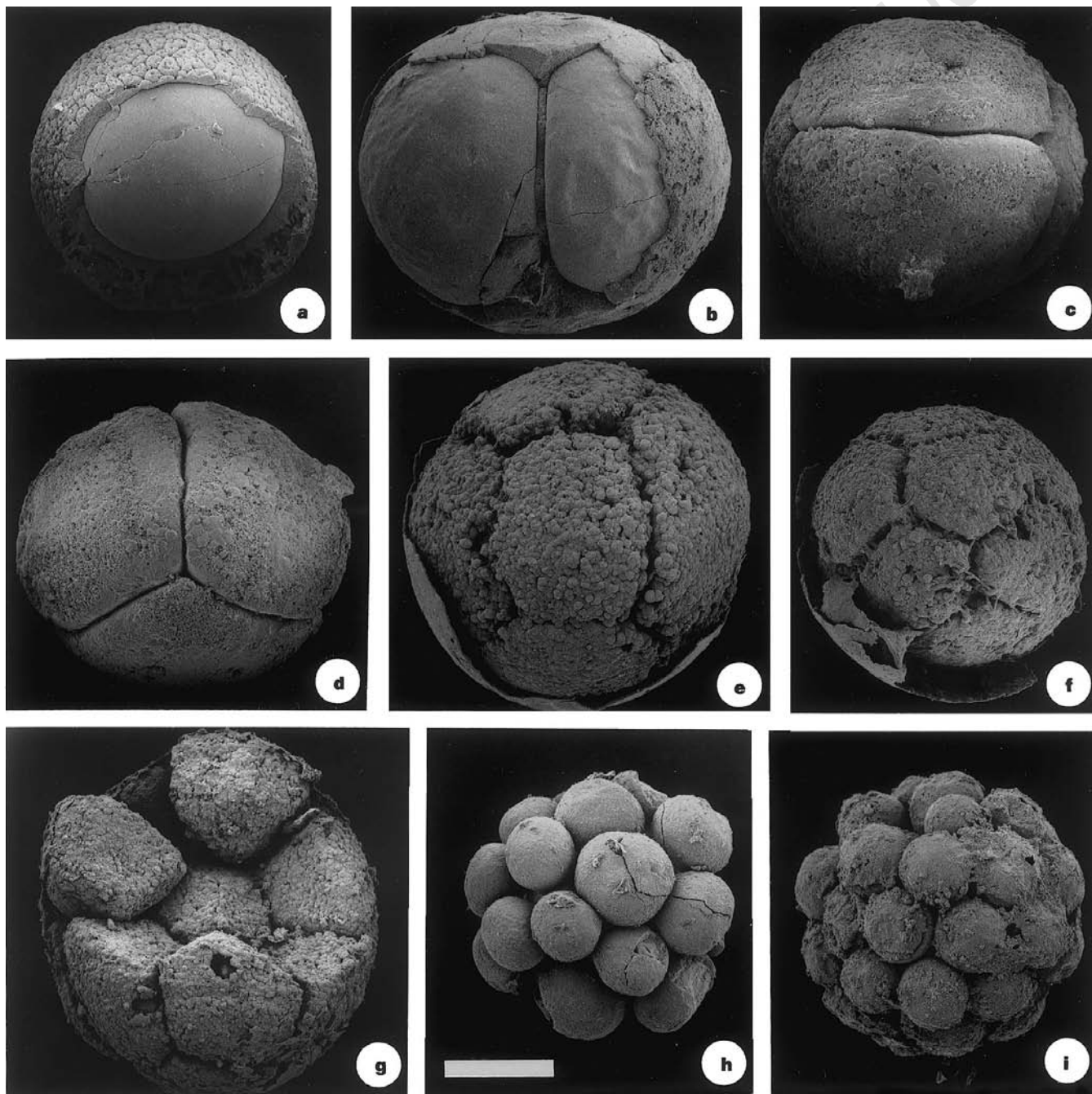


Figure 5 Fossil embryos preserving different stages of cleavage from the Doushantuo phosphorite. **a**, Fertilized (?) egg with thick membrane. **b**, Two-cell stage. **c**, **d**, Four-cell stage, **c** and **d** show different views of the same specimen, illustrating the tetrahedral geometry. **e**, Eight-cell stage. **f**, **g**, Later cleavage stages

showing faceted cell geometry and, in **g**, the three-dimensional distribution of cells. **h**, **i**, Multicellular structures that record later cleavage stages or, especially possible for **h**, colonial protists. Scale bar (in **h**): 200 μm for **a**, **e**, **f**, **g**, **h** and **i**; 150 μm for **b**; and 240 μm for **c** and **d**.

ported for at least some Doushantuo thalli by the differentiation of distinct medullary and cortical tissues coupled with reproductive structures similar to the carposporangia and spermatangia of living red algae (Fig. 4).

Compact thalli found abundantly in Doushantuo phosphorites^{13,24} contain ellipsoidal clusters of large (8–18 μm) dark cells distributed along thallus peripheries and occasionally protruding from the thallus surface (Fig. 4a–c). One or more layers of light-coloured, elongate cells (1–4 μm wide and up to 10 μm long) curve around and encompass the clusters. The cell clusters are interpreted as carposporangia and empty spheroidal regions that occupy a similar anatomical position are interpreted as empty conceptacles or cystocarps. Distinctive cell rows at the base of carposporangial clusters compare closely with the gonimoblast filaments that support carposporangial development in living red algae (Fig. 4a). Other thalli contain distinctive tissues of large, dark, oblong or rod-like cells arranged in parallel rows, with apical cells distinctly elongated and separated from subadjacent cell rows by transverse septa (Fig. 4d); these structures are interpreted as male reproductive tissues rather like the spermatangial sori of some extant rhodophytes (Fig. 4e).

Such features suggest affinities to the red-algal class Florideophyceae. Other Doushantuo fossils indicate that the class Bangiophyceae had also diversified markedly by the end of the Proterozoic eon: foliose thalli with regularly arranged cruciate cell tetrads (and, less commonly, octads) are indistinguishable from the carposporangial thalli of the extant bangiophyte alga *Porphyra* (Fig. 2c, d).

The oldest known fossil red algae are silicified *Bangia*-like filaments found in 1,200–900 Myr-old tidal flat carbonates from arctic Canada²⁹. Doushantuo fossils show that by the end of the Proterozoic eon the red algae were structurally complex and taxonomically diverse.

A comparable evolutionary pattern seems to characterize the photosynthetic stramenopiles. Compressed seaweeds in uppermost Doushantuo shales in the Yangtze Gorges area include a large population attributable to the Fucales, one of the most derived of all brown algal orders³⁰, whereas microscopic, coenocytic algae in the 1,000–900 Myr-old Lakhanda Group of eastern Siberia³¹ are indistinguishable from species of the extant xanthophyte *Vaucheria*. Similarly, the possible green algae in Doushantuo phosphorites and shales³² complement 750–700 Myr-old compression fossils from Spitsbergen that include populations similar to *Cladophora*³³, among the shallowest branching of all ulvophyte green algae. Thus, by the time large animals enter the fossil record, the three principal groups of multicellular algae had not only diverged from other protistan stocks but had evolved a surprising degree of the morphological complexity exhibited by living algae.

Neoproterozoic embryos

Since Haeckel³⁴ it has been broadly accepted that the earliest metazoans must have been microscopic organisms similar in development and morphology to the embryos or larvae of living animals. Davidson *et al.*³⁵ have expanded this view of animal 'prehistory' by hypothesizing that animals not only originated but underwent substantial early cladogenesis as minute, little-differentiated metazoans similar in form, function and ontogeny to the larvae of living invertebrates. Only later, after developmental toolkits and physiological tolerances had been well established, did macroscopic size and the adult body plans of extant phyla evolve within already discrete clades. Molecular clock estimates have suggested that the divergence of protostome and deuterostome animals took place as early as 1,000–1,300 Myr BP³⁶. These data are currently the subject of intensive methodological scrutiny, although Neoproterozoic algae provide support for the broad hypothesis that animals originated long before they became conspicuous elements of the geological record. As noted above, algal fossils document rapid eukaryotic divergence beginning at least 1,000 Myr BP, and

molecular phylogenies imply that the ancestors of animals diverged from other eukaryotes as part of this radiation³⁷. Three important groups of algae evolved multicellular forms early in their history, and it is reasonable to suggest that animals did so too. The conventional metazoan fossil record can be reconciled with molecular hypotheses by making the simple assumption that clade divergence and the evolution of large size and adult body plans within clades are distinct events separated by a considerable interval of time^{38,39}.

Palaeontological tests of such conjectures require Lagerstätten of the first order. Reports of phosphatized invertebrate embryos in Cambrian carbonates suggest that older phosphatic Lagerstätten may be the best places to search for early records of animal evolution^{40,41}. The exquisite preservation of Doushantuo algae specifically invites a search for microscopic animal remains, and this search has yielded positive results.

Xue *et al.*⁴² have described spheroidal microfossils containing geometrically arranged cells from Weng'an phosphorites, interpreting them as volvocacean green algae similar to extant *Pandorina*. Our collections, however, demonstrate that comparisons to volvocacean or any other algae are unlikely, given the size, geometry of cell division and structure of encompassing vesicles in the fossil population. Instead, we argue that these fossils are preserved embryos.

The specimens in our sample population are globular, measuring about 500 μm in diameter ($\bar{x} = 584 \mu\text{m}$, $s_x = 12 \mu\text{m}$, $n = 115$). Individuals contain one, two, four, eight or more closely packed internal bodies, the size and orientation of which suggest they are cells that underwent successive binary divisions with little or no intervening growth (Fig. 5). The diminishing size of internal bodies as their number multiplies is suggestive of early embryonic development. The geometric arrangement of the internal bodies is precise and is strikingly similar to the early cleavage stages of metazoan embryos⁴³. An ornamented external covering 10 μm thick encompasses single cells, which we interpret as resting zygotes within egg cases. We further interpret fossils containing two, four, eight or more internal bodies with polygonal or faceted geometries as cleaving embryos, with the internal bodies being blastomeres. Cross-sections (Fig. 5g) suggest that they are stereoblastulas. Developing embryos are not enveloped by a thick egg case, but instead are bounded by a thin wall rather like the zygotic membrane of living invertebrates. Judging from their egg size, we infer that the Doushantuo embryos underwent direct or lecithotrophic larval development.

At the four-cell stage, blastomeres are arranged in a modified tetrahedron, with opposite pairs of cells meeting across cross-furrows oriented perpendicular to one another at either end of the embryo (Fig. 5c). Structures similar to the cross-furrows of early embryos can also be seen at the 8- and 16-cell stages (Fig. 5e–f). Therefore, we interpret the Doushantuo fossils as holoblastic and equally cleaving stereoblastulas. Gastrulas or later developmental stages have not yet been identified.

The Doushantuo fossils could be broadly equivalent to blastaea or planuloids *sensu* Haeckel³⁴, embryos of microscopic animals as envisioned by Davidson *et al.*³⁵, or the earliest developmental stages of some as yet unrecognized macroscopic metazoan. Tetrahedral geometries comparable to those in the Doushantuo population are unusual in modern animal embryos, but not unknown^{43–46}. In fact, some living crustacean arthropods⁴³, have large eggs, direct development, a tetrahedral four-cell stage, and equal holoblastic cleavage in combination. Thus, the Doushantuo fossils are most probably bilaterian, and they may document cladogenesis within the Bilateria before the appearance of diverse Ediacaran assemblages. Given the age and architectural simplicity of these remains, however, phylogenetic interpretation is best approached with caution.

Despite the many uncertainties that surround the interpretation of the Doushantuo embryos, they provide the first direct geological evidence in support of the hypothesis that the main metazoan clades

diversified before the emergence of a conspicuous animal fossil record. More generally, they show that palaeontological investigation can tell us a great deal about the early history of metazoan evolution.

Conclusion

More than a century ago, Agassiz⁴⁷ recognized the 'three-fold parallelism' of patterns in ontogeny, systematics and biostratigraphy. The remarkable phosphatic thalli and embryos of the Doushantuo Formation show that unanticipated palaeontological observations, together with insights from molecular phylogeny and developmental genetics, can facilitate a modern integration of phylogeny, development and palaeontology that extends deeply into evolutionary history to address the early evolution of multicellular life. □

Received 29 December 1997; accepted 15 January 1998.

- Buss, L. W. *The Evolution of Individuality* (Princeton Univ. Press, NJ, 1987).
- Han, T.-M. & Runnegar, B. Megascopic eukaryotic algae from the 2.1 billion-year-old Negaunee Iron-Formation, Michigan. *Science* **257**, 232–235 (1992).
- Knoll, A. H. The early evolution of eukaryotes: a geological perspective. *Science* **256**, 622–627 (1992).
- Whittington, H. B. *The Burgess Shale* (Yale Univ. Press, New Haven, CT, 1985).
- Briggs, D. E. G., Erwin, D. H. & Collier, F. J. *The Fossils of the Burgess Shale* (Smithsonian Institution Press, Washington DC, 1994).
- Kidston, R. & Lang, W. H. On Old Red Sandstone plants showing structure from the Rhynie chert bed, Aberdeenshire. Parts I–IV. *Trans. R. Soc. Edinb.* **51**, 761–784; **52**, 603–627; **62**, 643–680; **52**, 831–854 (1917–1921).
- Remy, W., Gensel, P. J. & Hass, H. The gametophyte generation of some early Devonian land plants. *Int. J. Plant Sci.* **154**, 35–58 (1993).
- Zhao, Z., Xing, Y., Ma, G. & Chen, Y. *Biostratigraphy of the Yangtze Gorge Area, (1) Sinian* (Geological Publishing House, Beijing, 1985).
- Zhao, Z. *et al.* *The Sinian System of Hubei* (China University of Geosciences Press, Wuhan, 1988).
- Sun, W. Late Precambrian pennatulids (sea pens) from the eastern Yangtze Gorge, China: *Paracharnia* gen. nov. *Precambrian Res.* **31**, 361–375 (1986).
- Qian, Y., Chen, M. & Chen, Y. Hyolithids and other small shelly fossils from the Lower Cambrian Huangshandong Formation in the eastern part of the Yangtze Gorge. *Acta Palaeontol. Sinica* **18**(3), 207–232 (1979).
- Compston, W., Williams, I. S., Kirshvink, J. L., Zhang, Z. & Ma, G. Zircon U–Pb ages for the Early Cambrian time-scale. *J. Geol. Soc. Lond.* **149**, 171–184 (1992).
- Zhang, Y., Ying, L., Xiao, S. & Knoll, A. H. Permineralized fossils from the Terminal Proterozoic Doushantuo Formation, South China. *Paleontol. Soc. Mem.* (in the press).
- Grotzinger, J. P., Bowring, S. A., Saylor, B. Z. & Kaufman, A. J. Biostratigraphic and geochronologic constraints on early animal evolution. *Science* **270**, 598–604 (1995).
- Kaufman, A. J., Knoll, A. H. & Narbonne, G. M. Isotopes, ice ages, and terminal Proterozoic earth history. *Proc. Natl Acad. Sci. USA* **94**, 6600–6605 (1997).
- Benus, A. P. Sedimentologic context of a deep-water Ediacaran fauna (Mistaken Point Formation, Avalon zone, eastern Newfoundland). *Bull. N. Y. State Mus.* **463**, 8–9 (1988).
- Hofmann, H. J., Narbonne, G. M. & Aitken, J. D. Ediacaran remains from intertillite beds in northwestern Canada. *Geology* **18**, 1199–1202 (1990).
- Briggs, D. E. G., Kear, A. J., Martill, D. M. & Wilby, P. R. Phosphatization of soft-tissue in experiments and fossils. *J. Geol. Soc. Lond.* **150**, 1035–1038 (1993).
- Krajewski, K. P. *et al.* Biological processes and apatite formation in sedimentary environments. *Ecol. Geol. Helvet.* **87**, 701–745 (1994).
- Glenn, C. R. *et al.* Phosphorus and phosphorites: Sedimentology and environments of formation. *Ecol. Geol. Helvet.* **87**, 747–788 (1994).
- Föllmi, K. B. The phosphorus cycle, phosphogenesis and marine phosphate-rich deposits. *Earth Sci. Rev.* **40**, 55–124 (1996).
- Zhu, S. & Wang, Y. in *The Upper Precambrian and Sinian–Cambrian Boundary in Guizhou* (eds Wang, Y. *et al.*) 93–103 (People's Publishing House of Guizhou, Guiyang, 1984).
- Zhang, Y. Multicellular thallophytes with differentiated tissues from Late Proterozoic phosphate rocks of South China. *Lethaia* **22**, 113–132 (1989).
- Zhang, Y. & Yuan, X. New data on multicellular thallophytes and fragments of cellular tissues from Late Proterozoic phosphate rocks, South China. *Lethaia* **25**, 1–18 (1992).
- Awramik, S. M. *et al.* Prokaryotic and eukaryotic microfossils from a Proterozoic/Phanerozoic transition in China. *Nature* **315**, 655–658 (1985).
- Bold, H. C. & Wynne, M. J. *Introduction to the Algae* (Prentice-Hall, Englewood Cliffs, NJ, 1985).
- Fritsch, F. E. *The Structure and Reproduction of the Algae* Vols 1, 2 (Cambridge Univ. Press, 1965).
- van den Hoek, C., Mann, D. G. & Jahns, H. M. *Algae: An Introduction to Phycology* (Cambridge Univ. Press, 1995).
- Butterfield, N. J., Knoll, A. H. & Swett, K. A bangiophyte red alga from the Proterozoic of Arctic Canada. *Science* **250**, 104–107 (1990).
- Xiao, S., Knoll, A. H. & Yuan, X. Morphological reconstruction of *Miaohephyton bifurcatum*, a possible brown alga from the Terminal Proterozoic Doushantuo Formation, South China. *J. Paleontol.* (in the press).
- Hermann, T. N. *Organic World Billion Year Ago* (Nauka, Leningrad, 1990).
- Chen, M. & Xiao, Z. Discovery of the macrofossils in the Upper Sinian Doushantuo Formation at Miaohu, eastern Yangtze Gorges. *Sci. Geol. Sinica* **4**, 317–324 (1991).
- Butterfield, N. J., Knoll, A. H. & Swett, K. Paleobiology of the Neoproterozoic Svanbergfjellet Formation, Spitsbergen. *Fossils Strata* **34**, 1–84 (1994).
- Haeckel, E. The gastraea theory, the phylogenetic classification of the animal kingdom and the homology of the germ-lamellae. *Q. J. Microsc. Soc.* **14**, 142–165 (1874).
- Davidson, E. H., Peterson, K. J. & Cameron, R. A. Origin of bilaterian body plans: Evolution of developmental regulatory mechanisms. *Science* **270**, 1319–1325 (1995).
- Wray, G. A., Levinton, J. S. & Shapiro, L. H. Molecular evidence for deep Precambrian divergences among metazoan phyla. *Science* **274**, 568–573 (1996).
- Sogin, M. L. in *Early Life on Earth* (ed. Bengtson, S.) 181–192 (Columbia Univ. Press, NY, 1994).
- Vermeij, G. J. Animal Origins. *Science* **274**, 525–526 (1996).
- Fortey, R. A., Briggs, D. E. G. & Wills, M. A. The Cambrian evolutionary 'explosion': decoupling cladogenesis from morphological disparity. *Biol. J. Linn. Soc.* **57**, 13–33 (1996).
- Zhang, X. & Pratt, B. R. Middle Cambrian Arthropod embryos with blastomeres. *Science* **266**, 627–639 (1994).
- Bengtson, S. & Yue, Z. Fossilized metazoan embryos from the earliest Cambrian. *Science* **277**, 1645–1648 (1997).
- Xue, Y., Tang, T., Yu, C. & Zhou, C. Large Spheroidal Chlorophyta fossils from the Doushantuo Formation phosphoric sequence (late Sinian), central Guizhou, South China. *Acta Palaeontol. Sinica* **34**, 688–706 (1995).
- Kumé, M. & Dan, K. Invertebrate Embryology. (NOLIT, Belgrade, 1968).
- Anderson, D. T. *Embryology and Phylogeny in Annelids and Arthropods* (International Series of Monographs in Pure and Applied Biology, Vol. 50) (Pergamon, Oxford, 1973).
- Brusca, R. C. & Brusca, G. J. *Invertebrates* (Sinauer, Sunderland, MA, 1990).
- Nielsen, C. *Animal Evolution: Interrelationships of the Living Phyla* (Oxford Univ. Press, 1995).
- Agassiz, L. *Essay on Classification* (reprinted from *Contributions to the Natural History of the United States*, vol. 1, 1857 (Harvard Univ. Press, Cambridge, MA, 1962)).

Acknowledgements. We thank Y. Leiming for field assistance, E. Selig for technical help and S. Bengtson, D. McHugh, R. M. Woollacott, S. J. Gould, C. Nielsen, E. Ruppert, E. Davidson, A. Cameron and J. Henry for discussions and comments. This work was partly supported by grants from NSFC (to Y.Z.) and NSF (to A.H.K.).

Correspondence and requests for materials should be addressed to A.H.K.

YOURS TO HAVE AND TO HOLD BUT NOT TO COPY

The publication you are reading is protected by copyright law. Photocopying copyright material without permission is no different from stealing a magazine from a newsagent, only it doesn't seem like theft.

If you take photocopies from books, magazines and periodicals at work your employer should be licensed with CLA.

Make sure you are protected by a photocopying licence.



The Copyright Licensing Agency Limited
90 Tottenham Court Road, London W1P 0LP
Telephone: 0171 436 5931 Fax: 0171 436 3986

A stellar origin for the short-lived nuclides in the early Solar System

S. Sahijpal*, J. N. Goswami*, A. M. Davis†, L. Grossman†‡ & R. S. Lewis†

* Physical Research Laboratory, Ahmedabad – 380 009, India

† Enrico Fermi Institute, University of Chicago, 5640 South Ellis Avenue, Chicago, Illinois 60637, USA

‡ Department of the Geophysical Sciences, University of Chicago, 5734 South Ellis Avenue, Chicago, Illinois 60637, USA

Primitive meteorites contain isotopes that are the decay products of short-lived nuclides in the early Solar System^{1,2}. The relative abundances of these isotopes provide a means to determine timescales for the formation and accretion of primitive Solar System objects, the abundances of the parent nuclides being fixed when these objects solidified. The abundances can also be used to investigate the source of the nuclides (such as ⁴¹Ca, ²⁶Al, ⁶⁰Fe, ⁵³Mn and ¹⁰⁷Pd), although this is an area of controversy. The nuclides could have originated from a single stellar object^{2–6}, such as a nearby red-giant or a supernova. But observations of enhanced ion fluxes in a molecular cloud⁷ have led to other models^{8–10} in which these nuclides are formed by energetic particle irradiation of gas and dust in the protosolar molecular cloud; alternatively, irradiation by energetic particles from the active early Sun may have occurred within the solar nebula itself^{11–18}. Here we show that there is a correlation between the initial abundances of ⁴¹Ca

and ²⁶Al in samples of primitive meteorite (as inferred from their respective decay products, ⁴¹K and ²⁶Mg), implying a common origin for the short-lived nuclides. We can therefore rule out the mechanisms based on energetic particle irradiation, as they cannot produce simultaneously the inferred initial abundances of both nuclides. If, as our results suggest, a single stellar source is responsible for generating these nuclides, we can constrain to less than one million years the timescale for the collapse of the protosolar cloud to form the Sun.

We made the measurements of Mg and K isotope compositions in small domains (~10 μm) of refractory hibonite (CaAl_{12–24}Mg_xTi_xO₁₉) grains present within Ca–Al-rich inclusions (CAIs) or as isolated fragments in three primitive meteorites. The main objectives of this study were to look for ²⁶Mg and ⁴¹K excesses resulting from *in situ* decay of the two short-lived nuclides ⁴¹Ca (mean life, $\tau \approx 0.15$ Myr) and ²⁶Al ($\tau \approx 1$ Myr) present at the time of formation of the hibonites and to test for a possible correlation between the amounts of these nuclides present in the early Solar System. The analysed samples include fragments of individual crystals and grains present in hibonite–spinel-rich refractory spherules from the CM chondrite Murchison, and grains present in CAIs found in two CV3 chondrites, Allende and Efremovka. Seven of the analysed hibonites were hand-picked from an HF–HCl insoluble residue of a large piece of the Murchison meteorite. An additional hibonite fragment (SH-7) was hand-picked from an exposed surface of Murchison and another sample (BB-4; a spherical hibonite–spinel-rich object) was recovered from the dense fraction of the powder produced by freeze–thaw disaggregation of another piece of Murchison. Polished sections of three hibonite-bearing CAIs (HAL and USNM 3529-42 from Allende and E50 from Efremovka) were also included in our study. The Mg and K isotopic measurements were carried out at the Physical Research Laboratory, Ahmedabad, India, with a high-

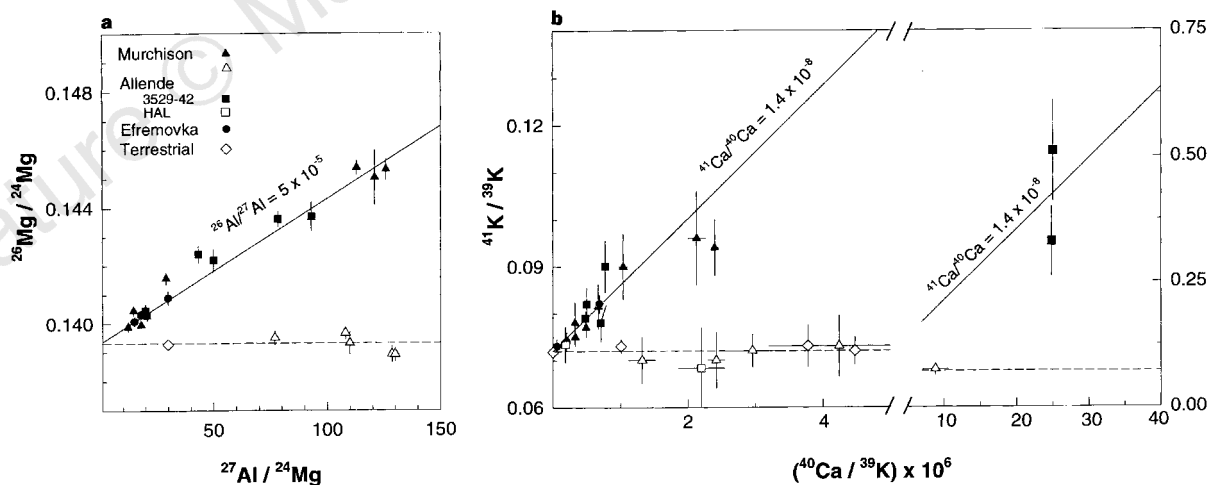


Figure 1 Magnesium and potassium isotopic compositions in hibonites from three primitive meteorites. Measured Mg and K isotopic ratios (²⁶Mg/²⁴Mg and ⁴¹K/³⁹K) in meteoritic and terrestrial hibonites plotted as a function of their ²⁷Al/²⁴Mg and ⁴⁰Ca/³⁹K ratios, respectively (error bars are 1σ). **a**, Magnesium isotopic data from 12 analyses of Murchison hibonites, together with three from the Efremovka CAI E40 and six from the Allende CAI 3529-42. The filled symbols represent hibonites from these meteorites with excess ²⁶Mg (²⁶Mg/²⁴Mg > 0.13932); the open symbols represent hibonites devoid of measurable ²⁶Mg excess. Data for terrestrial hibonite are consistent with the reference Mg isotope ratio (²⁶Mg/²⁴Mg = 0.13932; the dashed horizontal line). The solid line represents the expected evolution of a closed Mg–Al system with initial ²⁶Al/²⁷Al of 5×10^{-5} . Mg–Al isotopic data for the Allende (HAL) hibonite²¹ (not plotted) indicate high ²⁷Al/²⁴Mg ratio (>10⁴) with a very low initial ²⁶Al/²⁷Al of $\sim 5 \times 10^{-8}$. **b**, The K isotopic data; same symbols (filled or open) as in **a** are retained for representing

the analysed hibonites. The larger error bars result from the low count rates at masses 39 and 41 due to the low K content (<1 p.p.m.) in the hibonites. The solid lines represent the expected evolution of a closed K–Ca isotope system with initial ⁴¹Ca/⁴⁰Ca of 1.4×10^{-8} . The initial ⁴¹Ca/⁴⁰Ca values for the three data sets with ⁴¹K excess are $(1.41 \pm 0.58) \times 10^{-8}$ (Efremovka; three data points), $(1.09 \pm 0.34) \times 10^{-8}$ (Murchison hibonites; seven data points) and $(1.39 \pm 0.34) \times 10^{-8}$ (Allende 3529-42; six data points); errors are 2σ. Two analyses of Allende (HAL) hibonite did not reveal ⁴¹K excess (open squares). Another measurement of HAL yielded ³⁹K count rate below our counting system background (0.01 c s⁻¹); we infer an upper limit for the initial ⁴¹Ca/⁴⁰Ca from this analysis at 3×10^{-9} and a ⁴⁰Ca/³⁹K ratio of $> 2 \times 10^8$ (not plotted in the figure). The K isotopic ratios of the five Murchison hibonites devoid of ²⁶Mg excess and of terrestrial hibonite (Ca/K < 10³) and pyroxene (Ca/K > 10⁶) are consistent with the reference K isotopic ratio (⁴¹K/³⁹K = 0.072; the dashed horizontal lines).

mass-resolution Cameca ims-4f secondary ion mass spectrometer using procedures described in detail elsewhere^{19,20}. Measurement of the isotopic composition of Mg was followed by that of K at the same spot in each hibonite. In one case (HAL hibonite from Allende) we only measured the K isotope composition, as the Mg isotopic composition had been studied in detail earlier²¹.

In Fig. 1, we show the measured Mg and K isotopic compositions in the analysed hibonites. The details of the data reduction procedure are described elsewhere¹⁹. We carried out multiple analysis of hibonites in the Allende and Efremovka CAIs, and in two of the larger Murchison samples. Single analysis of both Mg and K isotopic compositions was carried out in the other smaller (< 50 µm) Murchison hibonites. The filled symbols in Fig. 1a represent hibonites from the three meteorites that have excess ²⁶Mg from *in situ* decay of ²⁶Al, the open symbols those devoid of any measurable excess. The measured ²⁶Mg/²⁴Mg ratio for terrestrial Madagascar hibonite is consistent with the reference value²² of 0.13932. The solid line represents the expected evolution for a closed Al–Mg isotope system with an initial ²⁶Al/²⁷Al ratio of 5×10^{-5} , the commonly accepted initial abundance ratio at the time of formation of most of the CAIs²³. In Fig. 1b, we show the results from the K isotope measurements. The data for terrestrial hibonite (and pyroxenes with high Ca/K ratios) are consistent with the reference ⁴¹K/³⁹K ratio of 0.072 (ref. 24) represented by the dashed horizontal lines. The solid lines represent the expected evolution for a closed K–Ca isotope system with an initial ⁴¹Ca/⁴⁰Ca of 1.4×10^{-8} , inferred from the data obtained from the study of the Efremovka CAIs^{19,25}.

The data presented in Fig. 1 show that hibonites with excess ²⁶Ni also have excess ⁴¹K and this correlation holds on a microscopic (~10 µm) scale. Further, hibonites having excess ²⁶Mg and ⁴¹K also have initial ²⁶Al/²⁷Al and ⁴¹Ca/⁴⁰Ca ratios close to their characteristic early Solar System values of 5×10^{-5} and 1.4×10^{-8} , respectively^{19,23}. We also note that when there is no excess ²⁶Mg corresponding to the decay of ²⁶Al (as in several Murchison hibonites), or when the inferred ²⁶Al/²⁷Al ratio is extremely low (for example in the Allende HAL hibonite²¹), we do not have any detectable ⁴¹K excess. We therefore conclude that the abundances of the two short-lived nuclides ²⁶Al and ⁴¹Ca in the early Solar System are intimately connected, indicating that they had a common origin. The absence of excess ²⁶Mg and ⁴¹K in some of the hibonites may indicate a coupled heterogeneous distribution of these two nuclides in the solar nebula, or secondary processing that may have caused redistribution or exchange of both Mg and K isotopes in these cases leading to the observed normal isotopic compositions. There is petrographic evidence²⁶ for secondary processing in the case of the Allende CAI HAL, notably in the hibonite. However, it is difficult to draw specific conclusions in the case of the Murchison hibonites devoid of ²⁶Mg and ⁴¹K excess, as the exact petrographic context of these hibonites, obtained by direct picking or acid dissolution of bulk meteorite samples, is unknown. The observed correlation between ²⁶Al and ⁴¹Ca would be further substantiated if samples with a lower initial ²⁶Al/²⁷Al and well behaved Al–Mg systematics could be found, where one would not expect excess ⁴¹K. For example, a sample with an initial ²⁶Al/²⁷Al value of $\sim 2.5 \times 10^{-5}$ would have an expected excess in ⁴¹K below our detection limit. Unfortunately, none of the samples analysed either by us or others satisfy the requirements of both a low initial ²⁶Al/²⁷Al and a high Ca/K ratio necessary for K isotopic study.

The observed correlations between the two short-lived nuclides ²⁶Al and ⁴¹Ca in refractory hibonite grains from primitive meteorites indicate that these two nuclides originated from a common source and followed the same pathways in the solar nebula before incorporation into these early Solar System solids. Although it is possible that both nuclides could have been produced during energetic particle irradiation either in a molecular cloud or in the early Solar System, we face the problem of matching the observed initial ²⁶Al/²⁷Al and ⁴¹Ca/⁴⁰Ca abundances of 5×10^{-5} and

1.4×10^{-8} , respectively. Extensive calculations that consider production of ⁴¹Ca, ²⁶Al, ⁵³Mn and ³⁶Cl by energetic particles within a molecular cloud complex^{9,10}, or by energetic particles emanating from an active early Sun and interacting with solar nebular gas^{11–13,16} or early Solar System solids^{13–15,18,27}, clearly show that production of enough ²⁶Al to match the observed initial ²⁶Al/²⁷Al ratio will result in a value of ⁴¹Ca/⁴⁰Ca more than an order of magnitude higher than observed. Overproduction of ⁵³Mn and ³⁶Cl by a similar magnitude is also predicted. One way to circumvent the problem of overproduction of the shorter-lived nuclide ⁴¹Ca ($\tau \approx 0.15$ Myr) will be to invoke a time interval of a few times 10^5 years, between the cessation of production by energetic particles and the formation of the analysed hibonites, so that the abundance of ⁴¹Ca goes down while that of ²⁶Al ($\tau \approx 1$ Myr) remains essentially unaffected. However, it is difficult to support such an *ad hoc* hypothesis and the problem of overproduction of the relatively longer-lived nuclide ⁵³Mn ($\tau = 5.3$ Myr) will remain.

On the other hand, specific models have been proposed to explain the observed initial ²⁶Al/²⁷Al and ⁴¹Ca/⁴⁰Ca ratios in early Solar System objects in terms of the production rates of these nuclides in different stellar sources such as a thermally pulsing asymptotic giant branch (TP-AGB) star, a massive supernova or a Wolf–Rayet star^{2–6,28}. The choice of the most probable stellar source is made difficult by the two free parameters involved in these models; the dilution of the freshly synthesized stellar material by material in the source region and later with protosolar cloud material, and the time interval between the injection of the freshly synthesized nuclides into the protosolar cloud and the formation of the early Solar System objects, such as the hibonites analysed in this work. However, if we consider the dilution factor to be similar for both ²⁶Al and ⁴¹Ca, a TP-AGB star seems to be a likely candidate¹⁹. A final resolution of the most plausible stellar source of these nuclides will require a better understanding of the interactions of these stellar objects with their surroundings^{2–6,28} as well as the dynamics of the eventual injection of freshly synthesized nuclides into the protosolar cloud, an event that could have triggered its collapse^{29,30}.

In conclusion, our data do not support the proposals that energetic particle interactions within the protosolar molecular cloud or later in the solar nebula led to the production of the short-lived nuclides ²⁶Al and ⁴¹Ca present in the early Solar System. Of course, we cannot completely rule out any contribution from this process; but if it is there, it must be at a much lower level than the contribution from the stellar source and may be ignored. Although the stellar source for the short-lived nuclides in the early Solar System cannot be identified unambiguously, a stellar origin for these nuclides, and particularly for ⁴¹Ca with a mean life of 0.15 Myr, constrains the timescale for the collapse of the protosolar cloud to form the Sun and some of the first Solar System objects to less than a million years^{4,5,19,28,31}. □

Received 29 May; accepted 14 November 1997.

1. Wasserburg, G. J. in *Protostars and Planets II* (eds Black, D. C. & Matthews, M. S.) 703–737 (Univ. Arizona Press, Tucson, 1985).
2. Cameron, A. G. W. in *Protostars and Planets III* (eds Levy, E. H. et al.) 47–73 (Univ. Arizona Press, Tucson, 1993).
3. Wasserburg, G. J., Busso, M., Gallino, R. & Raiteri, C. M. Asymptotic Giant Branch Stars as a source of short-lived radioactive nuclei in the solar nebula. *Astrophys. J.* **424**, 412–428 (1994).
4. Wasserburg, G. J., Gallino, R., Busso, M., Goswami, J. N. & Raiteri, C. M. Injection of freshly synthesized ⁴¹Ca in the early solar nebula by an asymptotic giant branch star. *Astrophys. J. Lett.* **440**, L101–L104 (1995).
5. Cameron, A. G. W., Höflich, P., Myers, P. C. & Clayton, D. D. Massive supernova, Orion gamma rays and the formation of the solar system. *Astrophys. J. Lett.* **447**, L53–L57 (1995).
6. Arnould, M., Paulus, G. & Meynet, G. Short-lived radionuclide production by non-exploding Wolf–Rayet stars. *Astron. Astrophys.* **321**, 452–465 (1997).
7. Bloemen, H. et al. COMPTEL observations of the Orion complex: Evidence for cosmic-ray induced gamma ray lines. *Astron. Astrophys.* **281**, L5–L8 (1994).
8. Clayton, D. D. Production of ²⁶Al and other extinct radionuclides by low-energy heavy cosmic rays in molecular clouds. *Nature* **368**, 222–224 (1994).
9. Clayton, D. D. & Jin, L. Gamma rays, cosmic rays, and extinct radioactivity in molecular cloud. *Astrophys. J.* **451**, 681–699 (1995).
10. Ramaty, R., Kozlovsky, B. & Lingenfelter, R. E. Light isotopes, extinct radioisotopes and gamma ray lines from low energy cosmic ray interactions. *Astrophys. J.* **456**, 525–540 (1996).
11. Heymann, D. & Diczkaniec, M. Early irradiation of matter in the solar system: Magnesium (proton, neutron) scheme. *Science* **191**, 79–81 (1976).

12. Clayton, D. D., Dwek, E. & Woosley, S. E. Isotopic anomalies and proton irradiation in the early solar system. *Astrophys. J.* **214**, 300–315 (1977).
13. Heymann, D., Dżiczkaniec, M., Walker, A., Huss, G. & Morgan, J. A. Effects of proton irradiation on a gas phase in which condensation take place. I. negative ^{26}Mg anomalies and ^{26}Al . *Astrophys. J.* **225**, 1030–1044 (1978).
14. Lee, T. A local proton irradiation model for isotopic anomalies in the solar system. *Astrophys. J.* **219**, 217–226 (1978).
15. Wasserburg, G. J. & Arnold, M. in *Lecture Notes in Physics 287, 4th Workshop on Nuclear Astrophysics* (eds Hillebrandt, W. et al.) 262–276 (Springer, Heidelberg, 1987).
16. Clayton, D. D. & Jin, L. A new interpretation of ^{26}Al in meteoritic inclusions. *Astrophys. J. Lett.* **451**, L87–L91 (1995).
17. Shu, F. H., Shang, H. & Lee, T. Toward an astrophysical theory of chondrites. *Science* **271**, 1545–1552 (1996).
18. Shu, F. H., Shang, H., Glassgold, E. & Lee, T. X-rays and fluctuating X-wind from protostars. *Science* **277**, 1475–1479 (1997).
19. Srinivasan, G., Sahijpal, S., Ulyanov, A. A. & Goswami, J. N. Ion microprobe studies of Efremovka CAIs: II. Potassium isotope composition and ^{41}Ca in the early solar system. *Geochim. Cosmochim. Acta* **60**, 1823–1835 (1996).
20. Goswami, J. N., Srinivasan, G. & Ulyanov, A. A. Ion microprobe studies of Efremovka CAIs: I. Magnesium isotope composition. *Geochim. Cosmochim. Acta* **58**, 431–447 (1994).
21. Fahey, A. J., Goswami, J. N., McKeegan, K. D. & Zinner, E. K. ^{26}Al , ^{244}Pu , ^{50}Ti , REE, and trace element abundances in hibonite grains from CM and CV meteorites. *Geochim. Cosmochim. Acta* **51**, 329–350 (1987).
22. Cantanzaro, E. J., Murphy, T. J., Garner, E. L. & Shields, W. R. Absolute isotopic abundance ratios and atomic weights of magnesium. *J. Res. NBS* **70A**, 453–458 (1966).
23. MacPherson, G. J., Davis, A. M. & Zinner, E. K. The distribution of aluminum-26 in the early solar system—A reappraisal. *Meteoritics* **30**, 365–386 (1995).
24. Garner, E. L., Murphy, T. J., Gramlich, J. W., Paulsen, P. J. & Barnes, I. L. Absolute isotopic abundance ratios and the atomic weight of a reference sample of potassium. *J. Res. NBS* **79A**, 713–725 (1975).
25. Srinivasan, G., Ulyanov, A. A. & Goswami, J. N. ^{41}Ca in the early solar system. *Astrophys. J. Lett.* **431**, L67–L70 (1994).
26. Allen, J. M., Grossman, L., Lee, T. & Wasserburg, G. J. Mineralogy and petrography of HAL, an isotopically-unusual Allende inclusion. *Geochim. Cosmochim. Acta* **44**, 685–699 (1980).
27. Goswami, J. N., Marhas, K. K. & Sahijpal, S. Production of short-lived nuclides by solar energetic particles in the early solar system (Abstr.). *Lunar Planet. Sci. XXVIII*, 439–440 (1997).
28. Arnold, M., Meynet, G. & Paulsen, P. J. in *Astrophysical Implications of the Laboratory Study of Presolar Materials* (eds Bernatowicz, T. & Zinner, E.) 179–202 (AIP Press, New York, 1997).
29. Boss, A. P. & Foster, P. N. in *Astrophysical Implications of the Laboratory Study of Presolar Materials* (eds Bernatowicz, T. & Zinner, E.) 649–664 (AIP Press, New York, 1997).
30. Cameron, A. G. W., Vanhala, H. & Höflich, P. in *Astrophysical Implications of the Laboratory Study of Presolar Materials* (eds Bernatowicz, T. & Zinner, E.) 665–696 (AIP Press, New York, 1997).
31. Cameron, A. G. W. The first ten million years in the solar nebula. *Meteoritics* **30**, 133–161 (1995).

Acknowledgements. We are grateful to G. J. Wasserburg, G. J. MacPherson and A. A. Ulyanov for providing the samples of Allende and Efremovka CAIs for our study. We also thank E. J. Olsen of the Field Museum of Natural History for specimens of Murchison and Allende from which SH-7, BB-4, HAL and the seven hand-picked grains were extracted. We appreciate comments received from A. G. W. Cameron and E. Zinner. We acknowledge support from the Department of Space, Government of India (S.S. and J.N.G.) and from the National Aeronautics and Space Administration (A.M.D., L.G. and R.S.L.). J.N.G. also acknowledges hospitality provided by the Lunar and Planetary Institute, Houston, during the preparation of this manuscript.

Correspondence and requests for materials should be addressed to J. N. Goswami (e-mail: goswami@prlernet.in).

Softening of nanocrystalline metals at very small grain sizes

Jakob Schiøtz, Francesco D. Di Tolla* & Karsten W. Jacobsen

Center for Atomic-scale Materials Physics and Department of Physics, Technical University of Denmark, DK2800 Lyngby, Denmark

Nanocrystalline solids, in which the grain size is in the nanometre range, often have technologically interesting properties such as increased hardness and ductility. Nanocrystalline metals can be produced in several ways, among the most common of which are high-pressure compaction of nanometre-sized clusters and high-energy ball-milling^{1–4}. The result is a polycrystalline metal with the grains randomly orientated. The hardness and yield stress of the material typically increase with decreasing grain size, a phenomenon known as the Hall–Petch effect^{5,6}. Here we present computer simulations of the deformation of nanocrystalline copper, which show a softening with grain size (a reverse Hall–Petch effect^{3,7}) for the smallest sizes. Most of the plastic deformation is due to a large number of small ‘sliding’ events of atomic planes at the grain boundaries, with only a minor part being caused by dislocation activity in the grains; the softening that we see at small grain sizes is therefore due to the larger fraction of

atoms at grain boundaries. This softening will ultimately impose a limit on how strong nanocrystalline metals may become.

To simulate the behaviour of nanocrystalline metals with the computer, we construct nanocrystalline ‘samples’ with structures similar to those observed experimentally: essentially equiaxed dislocation-free grains separated by narrow straight grain boundaries¹. Each sample contains 8–64 grains in a 10.6-nm cube of material, resulting in grain sizes from 3.3 to 6.6 nm. The grains are produced by a Voronoi construction⁸: a set of grain centres are chosen at random, and the part of space closer to a given centre than to any other centre is filled with atoms in an f.c.c. (face-centred cubic) lattice with a randomly selected crystallographic orientation. A typical ‘sample’ is shown in Fig. 1a. To mimic the system’s being deep within the bulk of a large sample, the system is replicated infinitely in all three spatial directions (periodic boundary conditions). The forces between the atoms are calculated with the effective-medium theory^{9,10}, which suitably describes many-atom interactions in metals. The metal chosen for these simulations is copper; very similar results were obtained with palladium. Before deforming the system we ‘anneal’ it by running a 50-ps molecular dynamics simulation at 300 K, allowing unfavourable configurations in the grain boundaries to relax. Doubling the duration of annealing does not have any significant effect, nor does an increase in the temperature to 600 K.

The main part of the simulation is a slow uniaxial deformation while minimizing the energy with respect to all atomic coordinates. The deformation is applied by expanding the simulation cell in one direction, while the size is allowed to relax in the two perpendicular directions.

The initial and final configurations of such a simulation with a total strain of 10% are shown in Fig. 1. We see how the grain boundaries have become thicker, indicating that significant activity has taken place there. In the grains a few stacking faults have appeared. They are the signature of dislocation activity within the grains.

To facilitate the analysis of the simulations, we identify which atoms are located at grain boundaries and which are inside the grains, by determining the local crystalline order^{11,12}. Atoms in local f.c.c. order are considered to be ‘inside’ the grains; atoms in local h.c.p. (hexagonal close-packed) order are classified as stacking faults. All other atoms are considered as belonging to the grain boundaries. Unlike conventional materials, where the volume occupied by the grain boundaries is very small, a significant fraction (30–50%) of the atoms are in the grain boundaries, in agreement with theoretical estimates².

As the deformation takes place, we calculate the average stress in the sample as a function of the amount of deformation. For each grain size we simulated the deformation of seven different initial configurations. Figure 2a shows the obtained average deformation curves. We see a linear elastic region with a Young’s modulus around 90–105 GPa (increasing with increasing grain size), compared with 124 GPa in macrocrystalline Cu (ref. 13). This is caused by the large fraction of atoms in the grain boundaries having a lower Young’s modulus^{14,15}. A similar reduction is seen in simulations where the nanocrystalline metal is grown from a molten phase¹⁶. The elastic region is followed by plastic yielding at around 1 GPa, and finally the plastic deformation saturates at a maximal flow stress around 3 GPa. The theoretical shear stress of a perfect single crystal is approximately 6 GPa for the potential used.

The main deformation mode is illustrated in Fig. 3b, where the relative motions of the atoms is shown. We see that most of the deformation occurs in the grain boundaries in the form of a large number of small sliding events, where only a few atoms (or sometimes a few tens of atoms) move with respect to each other. Occasionally a partial dislocation is nucleated at a grain boundary and moves through a grain. Such events are responsible for a minor part of the total deformation, but in the absence of diffusion they are

* Present address: SISSA, Via Beirut 2–4, I-34014 Grignano (TS), Italy.

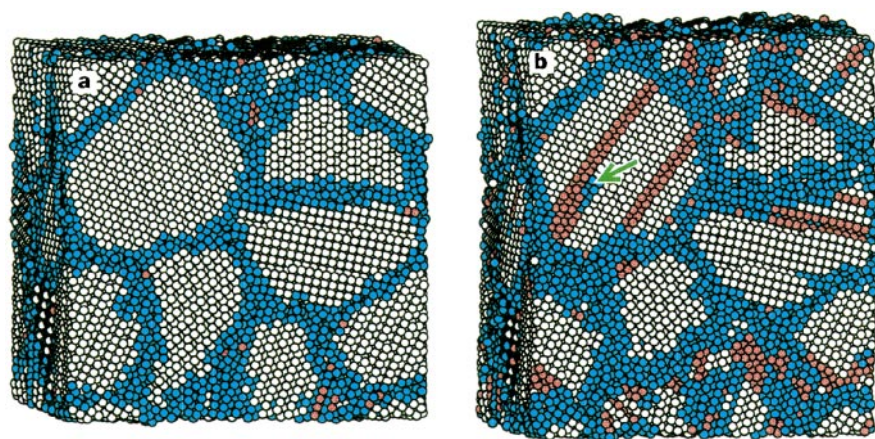


Figure 1 A simulated nanocrystalline copper sample before (a) and after (b) 10% deformation. The system contains 16 grains and $\sim 100,000$ atoms, giving an average grain diameter of 5.2 nm. Atoms in the grain boundaries are coloured blue, atoms at stacking faults are coloured red. We clearly see stacking faults left behind by partial dislocations that have run through the grains during the deformation processes. Such stacking faults would be removed if a second partial dislocation followed the path of the first, but this is not observed in the present simulations. In the left side of the system, a partial dislocation on its way through a grain is seen (green arrow in b).

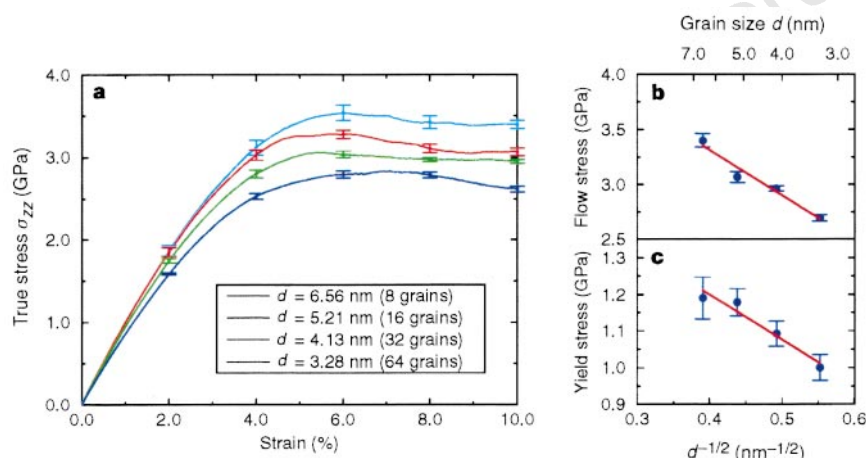


Figure 2 The effect of grain size on deformation. **a**, The average stress in the direction of the stretch (σ_{zz}) versus strain for each grain size. Each curve is the average over seven simulations. The curves show the response of the material to mechanical deformation. In the linear part of the curve (low strains) the deformation is mainly elastic; if the tensile load is removed, the material will return to the original configuration. As the deformation is increased, irreversible plastic deformation becomes important. For large deformations plastic pro-

cesses relieve the stress, and the curves level off. We see a clear grain-size dependence, which is summarized to the right. **b** and **c**, The maximal flow stress and the yield stress as a function of grain size. The yield stress decreases with decreasing grain size, resulting in a reverse Hall–Petch effect. (The maximal flow stress is the stress at the flat part of the stress–strain curves; the yield stress is defined as the stress where the strain departs 0.2% from linearity.)

required to allow for deformations of the grains, as they slide past each other. No dislocation motion is seen in Fig. 3, as none occurred at that time of the simulation. As the grain size is reduced a larger fraction of the atoms belongs to the grain boundaries, and grain-boundary sliding becomes easier. This leads to a softening of the material as the grain size is reduced (Fig. 2).

The observed deformation mode is in some ways similar to the manner in which grain boundaries carry most of the deformation in superplastic deformation^{17,18}. This is consistent with recent simulations of flow speed in nanocrystalline metals²⁵. However, in superplasticity the grain-boundary sliding is thermally activated, whereas here it occurs at zero temperature driven by the high stress.

In conventional metals an increase in hardness and yield strength with decreasing grain size is observed. This is called the Hall–Petch effect, and is generally considered to be caused by the grain boundaries, impeding the generation and/or motion of dislocations as the grains get smaller; this behaviour extends far into the nanocrystalline regime^{3,19}. The grain sizes in the present simulations correspond to the smallest grain sizes that can be obtained experimentally. In that regime the Hall–Petch effect is often seen to cease or even to reverse, but the results depend strongly on the sample history and on the method used to vary the grain size. Many mechanisms have been proposed for this reverse Hall–Petch effect: increased porosity at small grain sizes³, suppression of

dislocation pile-ups²⁰, dislocation motion through multiple grains²¹, sliding in the grain boundaries²², and enhanced diffusional creep in the grain boundaries⁷. Direct measurements of the creep rates seem to rule out the last mechanism^{19,20}, but otherwise no consensus has been reached. The present simulations indicate that the behaviour characteristic of a reverse Hall–Petch effect is possible even in the absence of porosity, and that it may be caused by sliding in the grain boundaries even in the absence of thermally activated processes. We cannot, however, see a cross-over from this ‘reverse’ behaviour to the normal Hall–Petch regime at larger grain sizes in our simulations, because they become too computationally expensive at these larger sizes. For the same reason, we cannot provide a direct comparison with the behaviour of the bulk metal.

A direct quantitative comparison between the simulations and the experimental results should be done with caution. The main difference between the simulated strain–stress curves and experimental curves is the level of the yield stress, which is approximately twice what is observed in experiments on low-porosity samples (400 MPa)²³. This value is, however, obtained for a grain size of around 40 nm; extrapolation to 7-nm grains gives a yield strength around 800 MPa (ref. 23), assuming that the Hall–Petch behaviour persists to these grain sizes. Experimentally produced nanocrystalline samples typically contain voids and surface defects, reducing the strength of the material. Surface defects alone have been shown

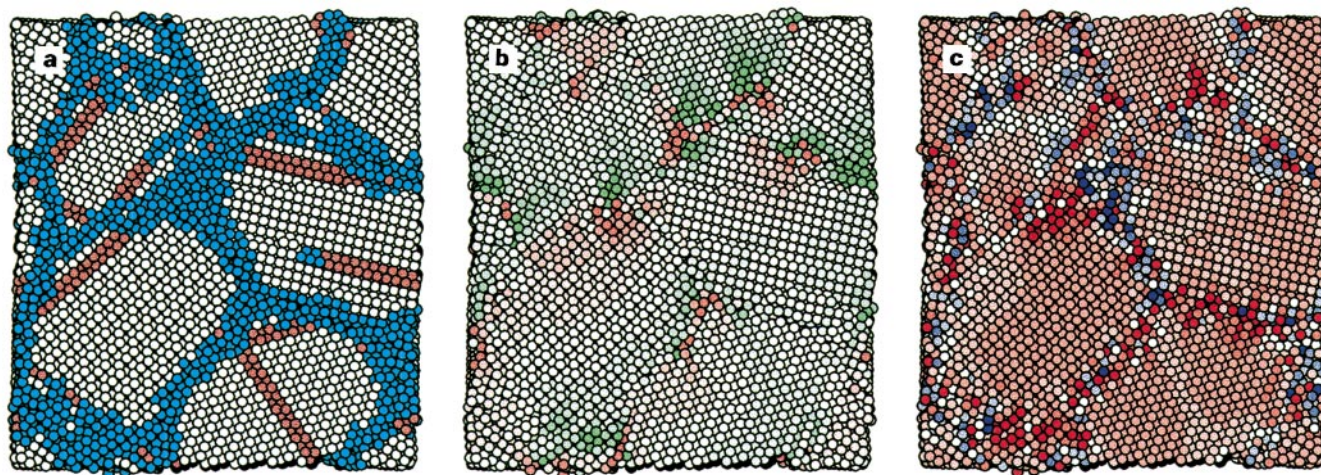


Figure 3 Snapshot of grain structure, displacements and stresses at 8% deformation. **a**, The position of the grain boundaries (blue) and stacking faults (red) at this point in the simulation. **b**, The relative motion of the atoms in the z direction (up, in the plane of the paper) during the preceding 0.4% deformation. The green atoms move up, red atoms move down. We see many small,

to be able to reduce the strength of nanocrystalline palladium by at least a factor of five^{19,24}.

Another difference is the absence of thermally activated processes in the simulations. These processes give rise to a strain-rate dependence of the mechanical properties, leading to higher yield stresses at higher strain rates where there is less time available for the activated processes to occur. Thermally activated processes are not included in the simulations because the energy minimization procedure quickly removes all thermal energy. No timescale or strain rate can be directly defined in the simulations, but in a sense the procedure corresponds to a slow strain at very low temperatures: thermally activated processes are excluded but the energy created by the work is carried away fast. The above-mentioned creep measurements^{19,20} indicate that diffusion does not play a major role during deformation.

During the later part of the simulated deformation larger average stresses build up within the grains than in the grain boundaries (10–20%), and larger stresses build up in the larger grains (see Fig. 3c). This results in a larger variation in the maximal flow stress than in the yield stress (Fig. 2b): when the grain size is increased the maximal flow stress increases, both because the stresses in the grains increase and because the number of atoms within the grains becomes a larger fraction of the total number of atoms. □

Received 23 June; accepted 29 October 1997.

1. Siegel, R. W. in *Encyclopedia of Applied Physics* Vol. 11, 173–199 (VCH, New York, 1994).
2. Siegel, R. W. What do we really know about the atomic-scale structure of nanophase materials? *J. Phys. Chem. Solids* **55**, 1097–1106 (1994).
3. Siegel, R. W. & Fougere, G. E. in *Nanophase Materials: Synthesis—Properties—Applications* (eds Hadjipanayis, G. C. & Siegel, R. W.) 233–261 (NATO-ASI Ser. E, Vol. 260, Kluwer, Dordrecht, 1994).
4. Gleiter, H. in *Mechanical Properties and Deformation Behavior of Materials Having Ultra-Fine Microstructures* (ed. Nastasi, M.) 3–35 (Kluwer, Dordrecht, 1993).
5. Hall, E. O. The deformation and ageing of mild steel: III Discussion of results. *Proc. Phys. Soc. Lond. B* **64**, 747–753 (1951).
6. Petch, N. J. The cleavage of polycrystals. *J. Iron Steel Inst.* **174**, 25–28 (1953).
7. Chokshi, A. H., Rosen, A., Karch, J. & Gleiter, H. On the validity of the Hall–Petch relationship in nanocrystalline materials. *Scripta Metall.* **23**, 1679–1684 (1989).
8. Voronoi, G. Z. *Reine Angew. Math.* **134**, 199 (1908).
9. Jacobsen, K. W., Nørskov, J. K. & Puska, M. J. Interatomic interactions in the effective-medium theory. *Phys. Rev. B* **35**, 7423–7442 (1987).
10. Jacobsen, K. W., Stoltze, P. & Nørskov, J. K. A semi-empirical effective medium theory for metals and alloys. *Surf. Sci.* **366**, 394–402 (1996).
11. Jónsson, H. & Andersen, H. C. Icosahedral ordering in the Lennard–Jones liquid and glass. *Phys. Rev. Lett.* **60**, 2295–2298 (1988).
12. Clarke, A. S. & Jónsson, H. Structural changes accompanying densification of random-sphere packings. *Phys. Rev. E* **47**, 3975–3984 (1993).
13. Gschneidner, K. A. Physical properties and interrelationships of metallic and semimetallic elements. *Solid State Phys.* **16**, 275–426 (1964).
14. Shen, T. D., Koch, C. C., Tsui, T. Y. & Pharr, G. M. On the elastic moduli of nanocrystalline Fe, Cu, Ni, and Cu–Ni alloys prepared by mechanical milling/alloying. *J. Mater. Res.* **10**, 2892–2896 (1995).

independent slip events in the grain boundaries; this is the main deformation mode. **c**, The stress field (the σ_{33} component) in the grains. Shades of red indicate tensile stress, shades of blue compressive stress (dark colours correspond to high stresses). The stress in the grain boundaries is seen to vary considerably on the atomic scale, and the average stress is ~10–20% lower than in the grains.

15. Kluge, M. D., Wolf, D., Lutsko, J. F. & Phillpot, S. R. Formalism for the calculation of local elastic constants at grain boundaries by means of atomistic simulation. *J. Appl. Phys.* **67**, 2370–2379 (1990).
16. Phillpot, S. R., Wolf, D. & Gleiter, H. Molecular-dynamics study of the synthesis and characterization of a fully dense, three-dimensional nanocrystalline material. *J. Appl. Phys.* **78**, 847–860 (1995).
17. Chokshi, A. H., Mukherjee, A. K. & Langdon, T. G. Superplasticity in advanced materials. *Mater. Sci. Eng. R* **10**, 237–274 (1993).
18. Ridley, N. (ed.) *Superplasticity: 60 years after Pearson* (Institute of Metals, London, 1995).
19. Nieman, G. W., Weertman, J. R. & Siegel, R. W. Mechanical behavior of nanocrystalline Cu and Pd. *J. Mater. Res.* **6**, 1012–1027 (1991).
20. Nieh, T. G. & Wadsworth, J. Hall–Petch relation in nanocrystalline solids. *Scripta Met. Mater.* **25**, 955–958 (1991).
21. Lian, J., Baudelet, B. & Nazarov, A. A. Model for the prediction of the mechanical behaviour of nanocrystalline materials. *Mater. Sci. Eng. A* **172**, 23–29 (1993).
22. Langdon, T. G. The significance of grain boundaries in the flow of polycrystalline materials. *Mater. Sci. Forum* **189–190**, 31–42 (1995).
23. Suryanarayana, R. et al. Mechanical properties of nanocrystalline copper produced by solution-phase synthesis. *J. Mater. Res.* **11**, 439–448 (1996).
24. Weertman, J. R. Hall–Petch strengthening in nanocrystalline metals. *Mater. Sci. Eng. A* **166**, 161–167 (1993).
25. Van Swygenhoven, H. & Caro, A. Plastic behavior of nanophase Ni: A molecular dynamics computer simulation. *Appl. Phys. Lett.* **71**, 1652–1654 (1997).

Acknowledgements. We thank J. K. Nørskov, T. Leffers, O. B. Pedersen, A. E. Carlsson and J. P. Sethna for discussions. The Center for Atomic-scale Materials Physics is sponsored by the Danish National Research Foundation.

Correspondence and requests for materials should be addressed to J.S. (e-mail: schiotz@fysik.dtu.dk).

Compound refractive optics for the imaging and focusing of low-energy neutrons

M. R. Eskildsen*, P. L. Gammel†, E. D. Isaacs†, C. Detlefs‡, K. Mortensen* & D. J. Bishop†

* Risø National Laboratory, PO Box 49, DK-4000 Roskilde, Denmark

† Bell Laboratories, Lucent Technologies, 700 Mountain Avenue, Murray Hill, New Jersey 07974, USA

‡ Ames Laboratory and Department of Physics and Astronomy, Iowa State University, Ames, Iowa 50011, USA

Low-energy neutrons are essential for the analysis and characterization of materials and magnetic structures. However, both continuous (reactor-based) and pulsed (spallation-based) sources of such neutrons suffer from low fluence. Steering and lensing devices could improve this situation dramatically, so increasing spatial resolution, detectable sample volume limits and even perhaps opening the way for the construction of a neutron

microscope. Neutron optics have to date exploited either Bragg diffraction^{1,2}, such as bent crystals, or reflection, as in mirror³ guides or a Kumakhov lens^{4,5}. Refractive optics remain an attractive alternative as they would permit full use of the beam cross-section, allow a compact and linear installation and, because of similarity to conventional optics, enable the use of commercial design and simulation tools. These advantages notwithstanding, single-element refractive optics have previously been considered impractical as they are too weakly focusing, too absorptive and too dispersive. Inspired by the recent demonstration⁶ of a compound refractive lens (CRL) for high-energy X-rays, we have designed, built and tested a prototype CRL for 9–20-Å neutrons by using readily available optical components: our CRL has gains greater than 15 and focal lengths of 1–6 m, well matched to small-angle neutron scattering.

Refractive optics for neutrons resemble those for visible light in that they rely on the difference in refractive index $1 - n = \delta$ between air and the lens material. For unpolarized neutrons of wavelength λ , $\delta = \rho b_c \lambda^2 / 2\pi$, where ρ is the density of atomic nuclei. The bound coherent scattering length, b_c , is generally positive⁷ and of the order of 5–10 fm. The refractive index is correspondingly less than one, making a converging lens concave rather than the familiar convex shape. For a single, symmetric biconcave lens the focal length is $f_0 = R/2\delta$ where R is the radius of curvature of the lens. This focal length is typically very long (~ 100 m) as $\delta < 1.5 \times 10^{-4}$. Single-element refractive neutron optics have been demonstrated⁸ in conjunction with bounds on the free neutron charge, but this realization is not useful in a more general experimental geometry. However, shorter focal lengths can be achieved with an array of thin lenses, a CRL. For N identical elements (lenses), the focal length is $f = f_0/N$. The large number of elements, however, can lead to substantial absorption, negating the gains due to focusing. The absorption cross-section $\sigma_a = \lambda b'_c/2$ is determined by the imaginary part of the bound coherent scattering length b'_c . For a practical lens

design, one wants to maximize the ratio of the focal strength to the absorption $\delta/2\sigma_a\rho = b_c\lambda^2/\pi\sigma_a$. In this regard, neutron optics enjoy an advantage over photon optics in that converging lenses are concave, meaning that the principal optic ray intercepts the minimum lens thickness.

In Table 1 we show elements and isotopes favourable for neutron optics ranked by figure of merit, which is dominated by σ_a . Alloys and crystals made of these elements can be used, with crystals preferred because of their reduced diffuse scattering at low wave-vector. For rapid and inexpensive prototyping, we used spherical biconcave MgF₂ lenses, available as stock items (International Scientific Products, Tarrytown, NY) intended for 4–5 μ m infrared optics. Following the successful trials reported here, we are negotiating with vendors to make improved optics from orientated graphite and BeO. In the future, one could contemplate cooled optics made from crystalline CO₂, which has the highest possible figure of merit and no incoherent scattering. Still, the system described here works well and is inexpensive and easy to use.

A photograph of our 30-element CRL is shown in Fig. 1. The overall transmission of this CRL was acceptable, 77% at neutron wavelength $\lambda = 9.5$ Å and 50% at 16 Å. The CRL was tested in the SANS beamline of the cold neutron guide hall at the Risø DR3 reactor, using the configuration shown in Fig. 1 inset. In order to quantify our optical system, we added the two circular apertures shown in the inset as A1 and A2. Either an area detector or scintillation camera recorded the image.

The basic focal properties of the CRL are shown in Fig. 2. At this wavelength, $\lambda = 14.1$ Å, focus occurs at 3.3 m. The deviations from the gaussian lens formula are expected from our simulations, and are largely a result of the extent of the CRL relative to its focal length. Similar results were obtained for other neutron wavelengths in the 9–20 Å range, with the expected λ^{-2} dependent focal length. Two factors contribute to the image size σ in the focal plane. First, the demagnified image of the aperture A1 of size Δ produces an image

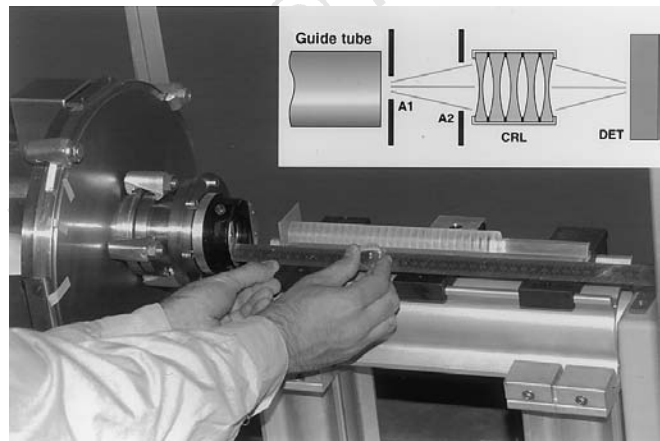


Figure 1 Photograph of the 30-element array of 25-mm-diameter symmetric spherical MgF₂ biconcave lenses, with one lens held aside for demonstration. Each lens had a radius of curvature of 25 mm, and centre thickness of 1 mm. For MgF₂, $\delta = 8.9 \times 10^{-5}$ at $\lambda = 10$ Å, corresponding to a focal length $f = 5.2$ m for the CRL. An incident aperture A1 was positioned 5.6 m from the CRL in an evacuated neutron guide. The guide is to the left in the photograph, and the lens aperture A2 is just to the left of the array. The neutron guide and mechanical velocity selector provide a neutron beam with wavelengths λ that range from 9 to 20 Å with a spread of $\Delta\lambda/\lambda = 9\%$. The angular beam divergence is very anisotropic: in the horizontal direction, it is a constant 20 arcmin and in the vertical, it is larger and wavelength-dependent at 12λ (Å) arcmin. Inset, a diagram of the lens test set-up. The collimating pinhole A1 is fixed 5.6 m from the compound refractive lens (CRL), and the 128×128 area detector (DET), also in an evacuated chamber, can be positioned 1.5–6.3 m from the CRL. The area detector pixel size was 5×5 mm.

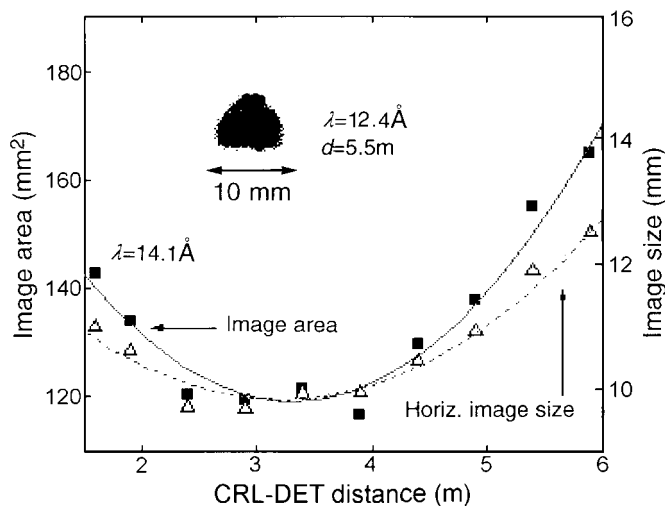


Figure 2 Focal properties of a 30-element lens array using 14.1-Å neutrons and a 2-mm pinhole A1 and 25-mm aperture A2. The horizontal image size (triangles) and total image area (squares), defined from the widths (full-width at half-maximum) of a two-dimensional gaussian fit are plotted versus CRL to detector distance. The image size is dominated by the detector resolution in this case, with a fit including all factors shown as the dashed line. Inset, a higher-resolution image of an 8.3-mm triangular aperture for $\lambda = 12.4$ Å and $d = 5.8$ m. This, lower-sensitivity, higher-resolution image shows two things: (1) the image is inverted as in a true optic, and (2) the image size is in agreement with the demagnification and smearing due to chromatic distortion.

size $\sigma = \Delta(3.3 \text{ m}/5.6 \text{ m}) \approx 1 \text{ mm}$ in both directions. Second is the chromatic aberration, yielding $\sigma = D(\Delta f/f) = 2D(\Delta\lambda/\lambda) = 4.5 \text{ mm}$, where D is the effective lens aperture and $\Delta\lambda/\lambda = 9\%$ is the wavelength spread of the neutrons. The spot size shown in Fig. 2 is much larger than either of these factors, because it is dominated by the coarse ($\sim 10 \text{ mm}$) and anisotropic resolution of the detector.

Table 1 Candidate elements for CRLs

Element/isotope	Atomic weight	b_c (fm)	σ_a (barns)	b_c/σ_a (fm $^{-1}$)
O	15.99	5.8	1.9×10^{-4}	310
C	12.01	6.6	3.5×10^{-3}	19
Be†	9.01	7.8	7.6×10^{-3}	10
Pb*	208.0	9.5	4.8×10^{-3}	8.0
F†	18.99	5.6	9.6×10^{-3}	5.8
Zr*	90.0	6.4	1.1×10^{-2}	5.3
Pb*	206.0	9.2	3.0×10^{-2}	3.1
Bi†	208.98	8.5	3.4×10^{-2}	2.5
H†	2.0	6.7	5.2×10^{-4}	2.1
Zr*	94.0	8.2	5.0×10^{-2}	1.6
Mg†	24.3	5.4	6.3×10^{-2}	0.86
Mo†	94.0	6.8	1.5×10^{-2}	0.85
Mo*	92.0	6.9	1.9×10^{-2}	0.68
Sr*	88.0	7.1	5.8×10^{-2}	0.43
N†	15.0	6.4	2.4×10^{-5}	0.34
Tl†	205.0	9.5	1.0×10^{-1}	0.24

A complete list of elements and isotopes useful for neutron lenses, with data for 2,200 m s $^{-1}$ (1.8 Å) thermal neutrons. Nuclei with $b_c > 5 \text{ fm}$, $\sigma_a < 0.1 \text{ barn}$ and natural abundance $> 5\%$ (for atomic weight > 40) and resultant ratio $b_c/\sigma_a > 0.1$ are included.

* For these elements, the final column has been adjusted for the relative natural abundance of different isotopes.

† Materials with an incoherent scattering length $b_i > 0.1 \text{ fm}$, which affects use with polarized neutrons, or in cases where diffuse scattering is important.

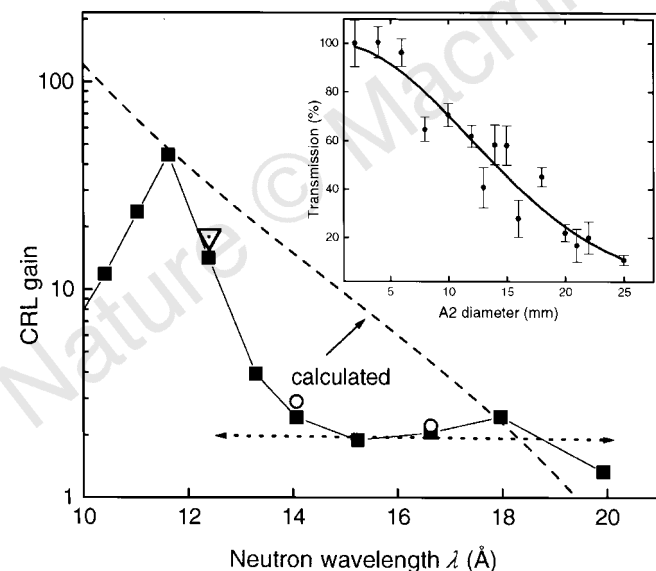


Figure 3 Dependence on neutron wavelength of the CRL gain (defined as intensity at focus for the CRL relative to pinhole collimation) with pinholes A1 of 2 mm (solid symbols) and 4 mm (open symbols). The higher gain at short wavelength is due to moving the focal point to large distances. The dotted line indicates that over more than an octave in wavelength, the gain > 2 . The gain estimated with the image of the inverted triangle aperture is shown as the triangle. The dashed line is a simple calculation of the gain, including aperture, absorption due to the centre thickness of the lens, chromatic distortion and demagnification. Inset, plot of total transmission of the lens versus size of aperture A2, used to determine the effective lens aperture at $\lambda = 14.1 \text{ Å}$. The transmission of the CRL, normalized to the same configuration with no lens, is plotted as a function of the size of the aperture A2. The reduced transmission for large A2 indicates that the effective lens aperture is beginning to determine the system aperture stop. Gaussian fits to the data give effective apertures of 15.7 mm at 9.5 Å, 11.9 mm at 14.1 Å.

In Fig. 2 inset, we show an actual image of a triangular aperture of 8.3 mm base length), taken with a lower-sensitivity, higher-resolution, scintillation camera. The inverted image is blurred by an amount consistent with chromatic aberration alone. This image is striking example of the focus and imaging properties of our CRL.

Two other factors can distort the focus, but are not significant here. Gravity^{9,10} leads to an effective index for the vacuum $\delta = gz/v_0^2$, where g is the acceleration due to gravity and v_0 is the neutron velocity. For 10 Å neutrons, $v_0 = 400 \text{ m s}^{-1}$ and $\delta = 6.1 \times 10^{-5} z$, where z (in m) is the height. An applied magnetic field B , proposed for soft neutron lenses, also gives an effective index change dependent on the neutron moment μ of $\delta = \pm \mu B/4\pi m v_0^2$.

The most important measure of a lens system is its gain, defined as the ratio of the intensity in a spot of the same area, replacing the CRL with a pinhole. In Fig. 3 we show the gain as a function of wavelength. Gains well in excess of 15 are easily achieved, and the CRL can be tailored to maximize the gain at any wavelength from 9 to 20 Å. The gain for the triangular aperture from Fig. 2 is also included as the triangular symbol. The dotted line indicates that broadband gains greater than 2 can also be achieved. This gain in intensity, similar to the gain from neutron supermirror guides, is quite rare and can make a significant difference in many types of experiments. The gain from focusing is reduced from the area of the lens image relative to pinhole optics by two attenuation factors, the absorption due to the centre thickness of the lens and the effective aperture. The concave shape of the individual lens elements creates an excess attenuation at the lens perimeter, which reduces the effective aperture from the physical diameter of the lens, as shown in Fig. 3 inset. A simple estimate of the gain, multiplying the image size by total attenuation, is shown by the dashed line in Fig. 3. The measured gain is below the calculation owing to detector smearing, not included in the gain data analysis. The peak in the gain is a simple artefact of the focal length exceeding 6 m, meaning that the detector can no longer be placed in the focal plane.

Simple compound refractive lenses are subject to large chromatic aberrations, as the focal length is proportional to λ^{-2} . For a wavelength spread $\Delta\lambda/\lambda = 9\%$, this gives $\Delta f/f = 18\%$. However, refractive optics can compensate for this by exploiting standard optical designs. For example¹¹, a separated doublet of a focusing and defocusing element can correct for longitudinal chromatic aberration. Similarly, a separated doublet consisting of two focusing elements corrects for lateral chromatic aberration. A CRL could also be combined with Fresnel optics¹², with their different chromaticity ($f_0 \propto \lambda^{-1}$). To fabricate these, and more complex optical systems, the most important factor will be the total transmission.

In general, CRLs provide a flexible, inexpensive method of steering and focusing low-energy, 9–20 Å, neutrons. Their similarity to visible optics allows the use of standard design techniques, including system optimization. One could even imagine using anomalous behaviour, such as large or negative coherent scattering lengths, associated with (n, γ) resonances in the range of thermal neutron energies, for the construction of exotic optics. Relative to mirrors, a CRL is more compact, requires lower surface quality, is easy to align and does not alter the beam path. Mirrors, of course, are inherently achromatic, although the critical angle for total reflection is both small and wavelength-dependent. Similar gains and image sizes have been reported with non-imaging micro-capillary arrays, known as Kumakhov lenses^{4,5}. In such an array, each capillary uses total external reflection to serve as a light pipe, and the entire array is tapered to produce a focal spot. However, relative to a Kumakhov lens, CRLs make more efficient use of the beam cross-section, and allow a more flexible design.

We have completed a first test of a CRL for low-energy neutrons, first proposed⁶ by Snigirev *et al.* We have achieved gains in excess of 15 relative to pinhole optics. The CRL, with 30 symmetric biconcave elements, had a wavelength-dependent focal length of 1–6 m, well matched to small-angle neutron scattering geometries. Similar to an

X-ray microprobe, our CRL could be used to guide and focus the neutron beam to explore local crystallography and magnetic domains. In addition, two other uses are possible. The first is to reduce neutron beam divergence and enhance the resolution limit of Bragg reflection rocking curves. The second possibility would be to magnify the pattern of scattered neutrons from a sample. Because area detectors for neutrons have a relatively low spatial resolution (5 mm), small-angle experiments use long (up to 40 m), expensive, evacuated paths to spread the diffraction pattern over the detector. Magnifying optics could be used to shorten this flight path. We are at present exploring new materials and achromatic lens designs. Such a system could provide significant new opportunities for the present generation of reactor-based neutron sources.

Received 13 June; accepted 4 November 1997.

1. Mikula, P., Kulde, J., Horalik, L., Chalupa, B. & Lukas, P. The spatial condensation of the neutron beam by an asymmetric diffraction in thermal-neutron monochromatization. *J. Appl. Crystallogr.* **19**, 324–330 (1986).
2. Majkrzak, C. F. & Wood, J. L. (eds) Neutron optical devices and applications *Proc. SPIE* **1738** (1992).
3. Sears, V. F. *Neutron Optics* (Oxford Univ. Press, 1989).
4. Kumakhov, M. A. & Sharov, V. A. A neutron lens. *Nature* **357**, 390–391 (1992).
5. Chen, H. *et al.* Guiding and focusing neutron beams using capillary optics. *Nature* **357**, 391–393 (1992).
6. Snigirev, A., Kohn, V., Snigireva, I. & Lengeler, B. A compound refractive lens for focusing high-energy X-rays. *Nature* **384**, 49–51 (1996).
7. Sears, V. F. Neutron scattering lengths and cross-sections. *Neutron News* **3**, 26–37 (1992).
8. Gähler, R., Kalus, J. & Mampe, W. An optical instrument for the search of a neutron charge. *J. Phys. E: Sci. Instrum.* **13**, 546–548 (1980).
9. Frank, A. I. Optics of ultracold neutrons and the neutron-microscope problem. *Sov. Phys. Usp.* **30**, 110–133 (1987).
10. Forgan, E. M. & Cubitt, R. Cancellation of gravity for slow neutrons in small angle scattering experiments (preprint, School of Physics and Astronomy, Univ. Birmingham, UK).
11. Jenkins, F. A. & White, H. E. *Fundamentals of Optics* (McGraw Hill, New York, 1976).
12. Klein, A. G. & Zeilinger, A. Wave optics with cold neutrons. *J. Phys. (Paris)* **45**, C3–239–C3–241 (1984).

Correspondence and requests for materials should be addressed to P.L.G. (e-mail: plg@lu.ac.uk).

Chiral discrimination by chemical force microscopy

Rachel McKendry*, Maria-Elena Theoclitou*, Trevor Rayment & Chris Abell

Department of Chemistry, Lensfield Road, Cambridge CB2 1EW, UK

* These authors contributed equally to this work.

Chirality is a fundamental aspect of chemical biology, and is of central importance in pharmacology. Consequently there is great interest in techniques for distinguishing between different chiral forms of a compound. Chemical force microscopy is a technique that combines chemical discrimination with atomic force microscopy by chemical derivatization of the scanning probe tip. It has been applied to the study of hydrophobic and hydrophilic interactions¹, the binding between biotin and streptavidin^{2,3} and between DNA bases⁴. Here we report on the use of chemical force microscopy to discriminate between chiral molecules. Using chiral molecules attached to the probe tip, we can distinguish the two enantiomers of mandelic acid arrayed on a surface, through differences in both the adhesion forces and the frictional forces measured by the probe.

The main potential advantage of chemical force microscopy (CFM) over other techniques which measure intermolecular forces, such as the surface force apparatus⁵, is its speed, combined with the ability to map the spatial arrangement of ligands on the surface¹. Measurements of the adhesion force between the tip and surface give information on the binding energy between the small number of molecules on the tip (typically 1–20) and those on the surface under the tip. This knowledge of the local adhesion on the surface may be combined with topographic and friction images to give functional maps of the surface, with the caveat that both the

local chemistry and the local mechanical properties due to differences in molecular packing contribute to the image contrast.

Because of the fundamental importance of chirality, we have sought to test whether CFM has sufficient sensitivity to discriminate between domains of different chirality on a surface probed with a chirally derivatized tip. The experiments are described schematically in Fig. 1. The tips were coated with an acylated phenylglycine **T**, modified with an alkanethiol, so that they could be attached to a gold-coated probe by chemisorption. This structure is one introduced by Pirkle⁶ as a stationary phase in chiral high-performance liquid chromatography (HPLC) columns. The combination of hydrogen-bonding interactions and π -stacking interactions in this molecule enables it to form often quite distinct transient diastereomeric complexes with the different enantiomers of a compound as they elute down the HPLC column. This leads to markedly different retention times and hence good chiral separations for carboxylic acids, alcohols and esters.

We chose to attach mandelic acid to the surface, as this acid is a small molecule with a single chiral centre, which can be attached by means of an alkanethiol to a gold surface in three different ways: as an ether (**M1**), through the aromatic ring (**M2**), or as an ester (**M3**). When the mandelic acid is bound as the ether **M1**, the three groups (A, B and C in Fig. 1b) presented to the chiral ligand are an acid, a phenyl ring and a hydrogen. These are each structurally dissimilar.

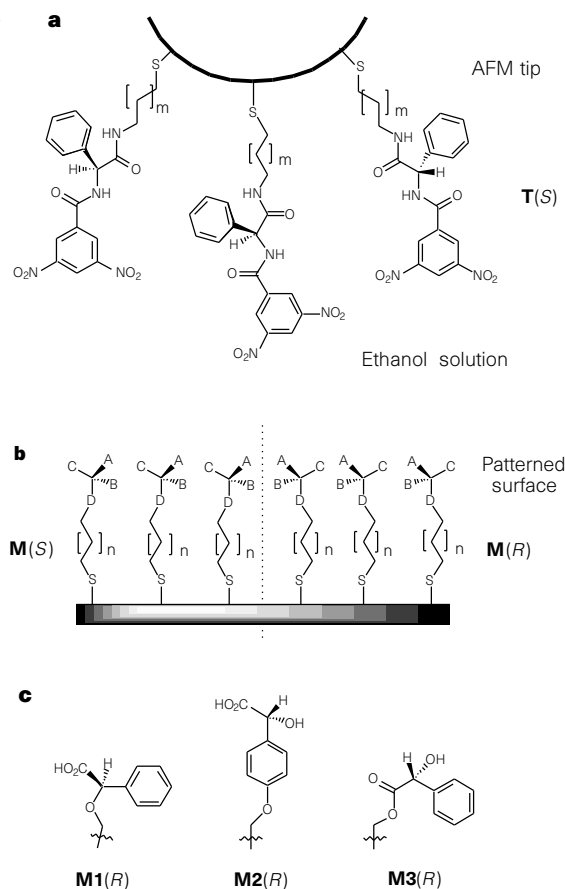


Figure 1 Schematic drawing of the experiment. **a**, The structure of the S-enantiomer of the chiral ligand **T** ($m = 10$) attached to the AFM tip, as described in the text. **b**, A surface coated with enantiomeric forms of a thiol **M** ($n = 9$) terminating in a chiral mandelic acid derivative. The boundary between general domains of opposite chirality are indicated by the vertical dashed line. **c**, Structure of the R-enantiomer of each of the mandelic acid thiols **M1**, **M2** and **M3**. The synthesis of this thiol **T** (R and S) and the differentially attached mandelic acids will be published elsewhere.

Furthermore they present hydrogen-bonding and π -stacking interactions complementary to the chiral ligand attached to the probe. We have made measurements of both adhesion and friction mapping on the systems described above. All the experiments were done under ethanol at 20 °C.

The results of measurement of the adhesion force for chiral tips and surfaces are shown in Fig. 2 and summarized in Table 1. There are 18 possible combinations of chiral surfaces and tips. For the sake of clarity full data are given for only a selection of these possibilities. The left-hand column of Fig. 2 shows distribution of adhesion or pull-off forces for a tip derivatized with the *S* enantiomer of **T** and mandelic acid bound as an ether to the surface (**M1**). The histogram shows that the interaction between the tip and the *R* form of the surface is more than twice that for the *S* surface. This result is mirrored by a stronger interaction between the *R* form of the tip and the *S* form of the surface. A racemic tip gives adhesion values intermediate to those of the *R* or *S* tip for a given surface. This might seem intuitively correct. However, a little further thought shows that this need not necessarily be the case, as the packing of molecules in racemic mixtures is invariably different from that of the pure enantiomers. Furthermore, phase separation and domain forma-

tion have been reported for 'racemic' surfaces^{7,8}. The disposition of *R* and *S* ligands on the racemic tip is unknown in these experiments.

The same pattern of adhesion forces is observed for **M2**, when the mandelic acid is bound through the aromatic ring. Now the three groups presented on the surface are an acid, an alcohol and a hydrogen group. Nevertheless the absolute values of adhesion forces are very similar to those for the mandelic acid derivative attached by an ether linkage. (There is some indication of enhanced discrimination for **M1** compared with **M2** but this observation lies on the limit of the statistical accuracy.) This is not what might be expected from chemical intuition, but very little is known of the packing or conformation of either the ligand or mandelic acid in these experiments.

Experiments with **M3**, mandelic acid attached through the carboxyl, were inconclusive. No discrimination of *R* and *S* surfaces could be detected through either friction mapping or measurements of adhesion forces. This may be because the carboxyl is not free to interact with the ligand and this might be the main interaction needed for chiral discrimination. We cannot, however, be dogmatic upon this point because we were unable to use the same experimental protocol in sample preparation. Further work is required to resolve this point.

For the friction measurements we prepared patterned surfaces. Figure 3 shows typical friction and topographic images obtained simultaneously from a pattern of alternate 2- μm -wide strips of the *R* and *S* mandelic acid derivatives of **M1** using a chirally functionalized tip, **T(S)**. Although there is clear evidence of a surface pattern in the friction map, there is no evidence for this in the raw

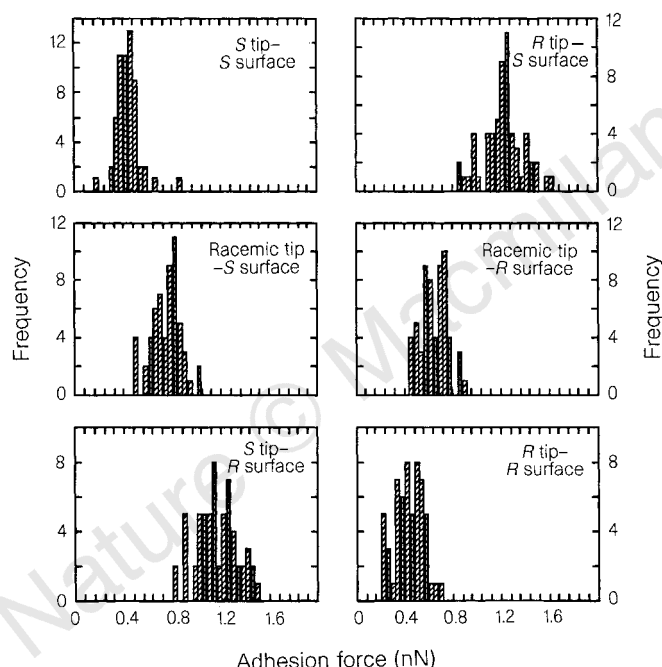


Figure 2 Histograms of the adhesion or 'pull-off' force. To minimize variations in data caused by use of different tips, measurements were made using a single tip on each of the six chiral surfaces. Sixty measurements of adhesion force were made on each surface in turn. The load-displacement curves were found to be of the standard form observed for studies of hydrophilic/hydrophobic interactions¹ and hence are not displayed here. The integrity of the tip was confirmed by repeating the measurements with **M1(S)** at end of the sequence. AFM tips were functionalized by first coating the tips (Park Scientific) with 100 nm gold by vacuum evaporation at room temperature. The cantilever was then immersed overnight at room temperature in a 1 mM solution of **T** in ethanol. The spring constants of the cantilevers were calculated from measurement of their resonant frequencies¹³ and gave a nominal value of 0.04 N m⁻¹. Samples for force spectroscopy were prepared by thermal evaporation of 100 nm of gold onto freshly cleaved mica substrates. The sample chamber was baked at 160 °C for at least 12 hours, the sample preheated to 340 °C, and evaporation conducted at pressures less than 10⁻⁸ mbar. The surfaces were then immersed overnight at room temperature in a 1 mM solution of **M** in ethanol. Before imaging, both the tip and surface were rinsed in ethanol and dried in nitrogen. Experiments were done under ethanol at 20 °C using a commercial AFM (East Coast Scientific).

Table 1 The adhesion or pull-off force for various combinations of chiral probes and surfaces

Surface	Tip		
	<i>S</i>	Racemic	<i>R</i>
M1(S)	0.4 ± 0.1	0.7 ± 0.1	1.2 ± 0.2
M1(R)	1.1 ± 0.1	0.6 ± 0.1	0.5 ± 0.1
M2(S)	0.5 ± 0.1	0.7 ± 0.1	1.1 ± 0.2
M2(R)	0.9 ± 0.2	0.7 ± 0.1	0.6 ± 0.1

Forces are in nN.

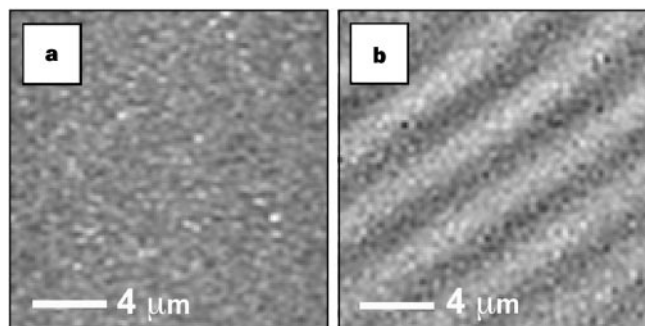


Figure 3 Maps of the surface. The topographic (a) and friction maps (b) were obtained simultaneously from a **M1** surface with a tip functionalized with **T(S)**. The scan rate was 5–30 $\mu\text{m s}^{-1}$ with a load of 10 nN. Patterns of alternating strips of *R*, *S* or racemic mandelic acid were produced by the microcontact printing scheme introduced by Kumar and Whitesides¹⁴. Samples were prepared by evaporation of 5 nm of chromium followed by 100 nm of gold onto a silicon surface at room temperature. An initial pattern of 2- μm -wide strips was printed from a stamp soaked in *R* mandelic acid. The integrity of the initial pattern was confirmed by scanning electron microscopy. The 'bare' strips were then functionalized by soaking the patterned substrate in a 1 mM ethanolic solution of the *S* mandelic acid overnight.

topographic images or in the 'error signal' maps which can be more sensitive to small topographic features. We have made extensive searches with a considerable number of samples prepared on different occasions to confirm that the lack of visible topographic detail is characteristic of these patterned surfaces. In addition we took care to ensure that the stripes observed in the friction map are genuine. Using tips functionalized with a racemic mixture of the chiral ligand **T**, we searched for evidence of contrast in both the topographic and friction images (experiments were performed with six racemic tips). No contrast was found. This null result for a racemic tip is consistent with the measurements of adhesion, and it implies that there are equal numbers of *R* and *S* molecules of the resin on the tip within the contact region.

The results above show that chemical force microscopy is sufficiently sensitive to permit discrimination between enantiomers of simple chiral molecules. For the systems used here, there is a difference of essentially 100% in the pull-off force for different enantiomers, and this difference in adhesion displays itself in frictional contrast on enantiomerically patterned surfaces. The chiral discrimination observed in these experiments must arise solely from the different diastereomeric complexes transiently formed between the chiral molecule on the tip and either the *R* or *S* mandelic acid on the surface. (The retention times for homologues of our mandelic acid derivatives, on a HPLC column containing the resin used here, give the same pattern as for the adhesion forces.) In other CFM studies that purportedly demonstrate chemical discrimination there is topographic and frictional contrast arising from variation of the mechanical properties of the film caused by variation in the intermolecular packing in the self-assembled monolayer for differently functionalized surfaces. We have eliminated this artefact from our experiments by use of surface patterns that differ only in their handedness. The packing and interfacial free energies are the same for the enantiomers in each pair, as confirmed by measurements of contact angles.

Now that chiral discrimination through CFM has been demonstrated, the next step is to gain a quantitative understanding of the results at a molecular level. We are investigating the packing and orientation of the functional groups of the self-assembled monolayers by means of Fourier-transform infrared spectroscopy. Coupled with molecular modelling this should improve understanding of the differences between the derivatives **M1**, **M2** and **M3**. Explanation of the adhesion energies requires an appropriate mechanical model for the interactions between the tip and surface. Work has shown that the JKR model^{5,9} of mechanical contacts is a good starting point for quantitative analysis of CFM¹⁰, and therefore to test this model's applicability, we are measuring the tip radius by scanning and transmission electron microscopy, and the surface energies by contact angle measurements. In these experiments we have measured both adhesion and friction, and although there is a link between these, its nature is one of the outstanding problems in tribology¹¹. From a qualitative chemical viewpoint it is clear that in these experiments the variation of friction between enantiomers must arise from the breaking of different diastereomeric complexes transiently formed in the region between chiral molecules on the tip and either the *R* or *S* mandelic acid on the surface. More experiments are being undertaken to improve the resolution and accuracy of the friction force maps¹².

The observation of chiral discrimination, which is fundamental to drug–target interactions, demonstrates that CFM can be used in the sophisticated way necessary to look for differential binding interactions between arrayed libraries of small molecules or oligonucleotides. In addition, experiments such as these may permit a step forward in our understanding of the subtle relationship between friction and adhesion, as they allow the interaction between the tip and surface to be changed without alteration of the mechanical properties of the surface or the wetting energies of tip or surface. □

Received 20 August; accepted 24 November 1997

1. Frisbie, C. D., Rozsnyai, L. F., Noy, A., Wrighton, M. S. & Lieber, C. M. Functional group imaging by chemical force microscopy. *Science* **265**, 2071–2074 (1994).
2. Lee, G. U., Kidwell, D. A. & Colton, R. J. Sensing discrete streptavidin–biotin interactions with atomic force microscopy. *Langmuir* **10**, 354–357 (1994).
3. Ludwig, M., Dettmann, W. & Gaub, H. E. Atomic force microscope imaging contrast based on molecular recognition. *Biophys. J.* **72**, 445–448 (1997).
4. Boland, T. & Ratner, B. D. Direct measurement of hydrogen bonding in DNA nucleotide bases by atomic force microscopy. *Proc. Natl Acad. Sci. USA* **92**, 5297–5301 (1995).
5. Israelachvili, J. *Intermolecular and Surface Forces* (Academic, New York, 1992).
6. Pirkle, W. H., Finn, J. M., Schreiner, J. L. & Hamper, B. C. A widely useful chiral stationary phase for the high-performance liquid chromatography separation of enantiomers. *J. Am. Chem. Soc.* **103**, 3964–3966 (1981).
7. Weis, R. M. & McConnell, H. M. Two-dimensional chiral crystals of phospholipid. *Nature* **310**, 47–49 (1984).
8. Stevens, E., Dyer, D. J. & Walba, D. M. Direct observation of enantiomorphous monolayer crystals from enantiomers by scanning tunneling microscopy. *Angew. Chem. Int. Edn Engl.* **35**, 900–901 (1996).
9. Johnson, K. L. Continuum mechanics of adhesion and friction. *Langmuir* **12**, 4510–4513 (1996).
10. Noy, A., Frisbie, C. D., Rozsnyai, L. F., Wrighton, M. S., Lieber, C. M. Chemical Force Microscopy: exploiting chemically-modified tips to quantify adhesion, friction and functional group distribution in molecular assemblies. *J. Am. Chem. Soc.* **117**, 7943–7951 (1995).
11. Yoshizawa, H., Chen, Y.-L. & Israelachvili, J. Fundamental mechanism of interfacial friction. I. Relation between adhesion and friction. *J. Phys. Chem.* **97**, 4128–4140 (1993).
12. Lantz, A. M., O'Shea, S. J., Welland, M. E. & Johnson, K. L. Atomic force microscope study of contact area and friction on NbSe₂. *Phys. Rev. B* **55**, 10776–10785 (1997).
13. Sader, J. E., Larson, I., Mulvaney, P. & White, L. R. Method for the calibration of atomic force microscope cantilevers. *Rev. Sci. Instrum.* **66**, 3789–3798 (1995).
14. Kumar, A. & Whitesides, G. M. Features of gold having micrometer to centimeter dimensions can be formed with an elastomeric stamp and an alkanethiol 'ink' followed by chemical etching. *Appl. Phys. Lett.* **63**, 2002–2004 (1993).

Acknowledgements. We thank S. O'Shea for discussions, G. Whitesides (Harvard University) and H. Bieybuck (IBM Zurich) for providing microcontact stamps, and R. Mellor for programming. We acknowledge financial support from Zeneca for a studentship (R.A.M.), and EPSRC for a postdoctoral fellowship (M.-E.T.).

Correspondence and requests for materials should be addressed to T.R. (e-mail: tr22@cam.ac.uk).

Spontaneous assembly of marine dissolved organic matter into polymer gels

Wei-Chun Chin, Mónica V. Orellana & Pedro Verdugo

Department of Bioengineering, University of Washington, Box 357962, Seattle, Washington 98195, USA

A large pool of organic carbon resides in the world's oceans in the form of dissolved organic matter (DOM)^{1,2}. DOM is operationally defined as the fraction of organic matter that passes through a filter with a given pore size (which can range from less than 0.1 µm to 0.46 µm). This fraction has a longer oceanic residence time—and is generally less biodegradable—than particulate organic matter (POM)^{1–4}. Processes transforming DOM into POM are therefore crucial for our understanding of the cycling of organic material in the oceans^{1–4}. The aggregation of marine colloids, which constitute 10–40% of DOM^{2,3,5}, is thought to be an important step in the transformation of DOM into POM³. It has been suggested that colloids, as well as transparent exopolymer particles and large aggregates ('marine snow') can be viewed as polymer gels^{6–8}. Whether free DOM polymers can indeed spontaneously assemble to form polymer gels has, however, not yet been shown. Here we present experimental observations that demonstrate that marine polymer gels can assemble from free DOM polymers, and that their formation mechanism, physical characteristics and mineralization can be understood in terms of polymer gel theory^{9–11}. The principles and methods of polymer gel physics thus have the potential to provide profound new insights into the processes controlling the exchange between the DOM and POM pools and the cycling of marine organic matter.

Polymer gels are formed by a three-dimensional polymer network and a solvent which in the case of hydrogels is water. Although the

solvent prevents the collapse of the network, the network entraps the solvent, creating a micro-environment that is in thermodynamic equilibrium with the surrounding media. The polymer chains that form the network are interconnected by chemical or physical crosslinks and thus stay in a statistically stable neighbourhood. Depending on the characteristics of the polymer chains (such as polyelectrolytic properties, degree of hydrophobicity, length, linear or branched chains), and the dielectric properties of the solvent, the polymers in the gel's matrix can interact strongly with each other, with the solvent, or with smaller solutes. These interactions determine the gel's physical properties. The study of polymer gels is an active area of research in modern physics, and here we use the insights gained in such studies to investigate the processes affecting marine organic matter.

The kinetics of DOM polymer assembly were monitored by measuring particle size as a function of time using dynamic laser scattering (DLS) and flow cytometry (Fig. 1a). Results show that DOM polymers in 0.22 μm -filtered sea water from Puget Sound, the North Pacific Ocean and the Arctic Ocean, can spontaneously assemble, forming polymer gels ranging from colloidal to micrometre size. In these experiments sea water suffered no processing other than gentle gravity filtration. DLS shows that fresh filtrate contains a polydisperse set of polymers with sizes ranging from 2 to 200 nm. About 30 minutes after filtration, DLS reveals the presence of a polydisperse collection of assembled polymer gels (POM) ranging in size from 200 nm to 1 μm . The assembly process is nonlinear and the increase in particle size follows a sigmoidal course resembling second-order kinetics. Independent measurements of particle size as a function of time, performed in replicate samples using flow cytometry, reveal a similar pattern of assembly. Both methods show that microgels continue growing for the next 50 h, reaching an average equilibrium size of $\sim 5 \mu\text{m}$ in 50–83 h. These techniques provide independent evidence that DOM polymers can undergo rapid spontaneous assembly into POM particles which, as shown below, have the typical features of polymer gels.

A control experiment showed that the assembly of microgels in

sea water filtered either once or twice (with a period of 8 days between the two filtrations) follows, within experimental error, the same time course. These results, together with environmental scanning electron microscopy (ESEM) observations that microgels are retained intact by the filter (Fig. 3, inset a), confirm that gravity filtration gently removes the microgels. As the assembly kinetics are concentration dependent, the observation also suggests that only a small fraction of the free DOM polymers forms microgels. This is in agreement with recent studies^{2,12} which suggest that only $\sim 10\%$ of DOM is transformed into POM.

The equilibrium particle size of the assembled gels depends on the total amount of available free polymers. Because there is a finite amount of polymers in the 10-ml samples, the average equilibrium size of microgels is only $\sim 5 \mu\text{m}$. But in the ocean, with a virtually infinite source of free polymers, the gels could potentially reach larger sizes.

Flow cytometry reveals that while the concentration of microgels decreases exponentially, their average size undergoes a corresponding increase (Fig. 2). These measurements, together with the sigmoidal time course of increase in particle size (Fig. 1a), imply that the formation and growth of marine microgels result both from the continuous assembly of free polymers into microgels and from secondary aggregation and annealing of small assembled gels. This is further supported by the observation that in each of the three replicates tested the total volume of gel remains constant over the time course of assembly.

Divalent bonds are known to stabilize the matrix of polyanionic gels¹³. In the case of marine hydrogels, chelation of Ca^{2+} and Mg^{2+} by addition of 10 mM ethylenediamine-tetra-acetic acid (EDTA) to the sea water inhibits polymer assembly (Fig. 1a). Furthermore, EDTA can readily disperse marine microgels, releasing free polymers of about 1 to 2.5 nm in size. This size range is known to be the main fraction of DOM^{2,14}. Dispersion induced by EDTA provides evidence that the polymer matrix of these gels is an ionic-bond-stabilized tangled network, rather than a covalently crosslinked network. The presence of bound Ca^{2+} inside the microgels was

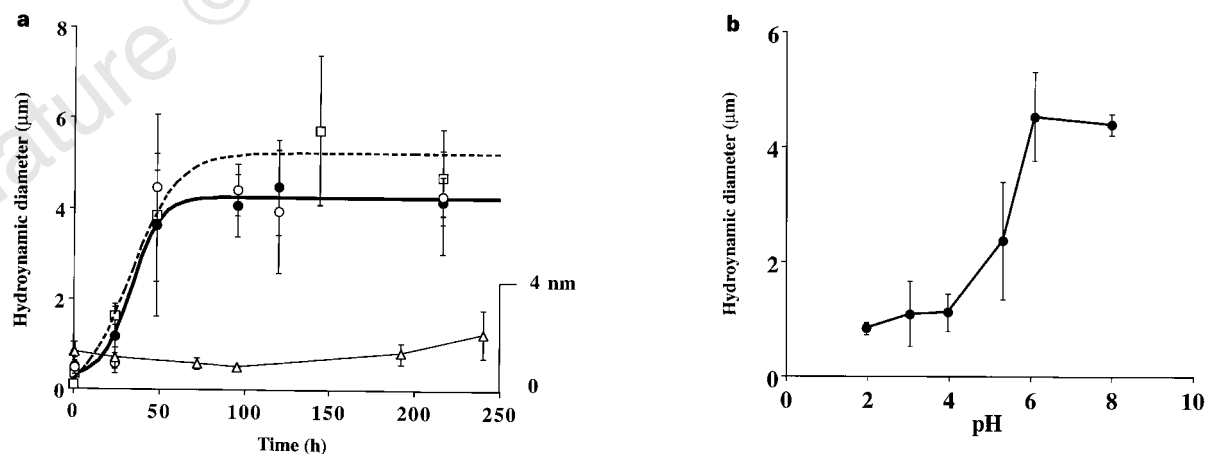


Figure 1 Polymer gel size as a function of assembly time and pH. **a**, The assembly of polymer gels was monitored by measuring particle size by both homodyne dynamic laser scattering (filled circles, solid line), and by flow cytometry (empty squares, dashed line). Parallel measurements were conducted in sea water containing 0.02% sodium azide (empty circles) to inhibit microbial activity. Control experiments in which the polymer assembly was inhibited by chelating Ca^{2+} and Mg^{2+} by addition of 10 mM EDTA to the sea water gave an average size of 1–2.5 nm, regardless of the time of observation (empty triangles, solid line). Measurements of microgel size by flow cytometry and DLS show a similar time course of assembly. We note that, depending on assembly time, the size of assembled polymer gels ranges from colloidal, submicrometre size, to several micrometres. Each point corresponds to the average $\pm$ s.d. of five measurements

made on three replicate samples of sea water. **b**, Experiment showing that marine polymer gels can undergo a fast (< 1 min) swelling/condensation transition from hydrated to condensed phase by changing the pH of the sea water. The size of the gels was monitored by DLS, and also verified by optical microscopy (data not shown). The swelling/condensation transition is reversible, and has a steep sigmoidal change in the volume of the gels that reflects the characteristic high apparent cooperativity of this phenomenon. Each point corresponds to the average $\pm$ s.d. of five measurements made on three replicate samples of sea water. Measurements by flow cytometry performed in the same samples showed no statistical difference in the concentration of microgels in swollen or condensed phase.

revealed by chlortetracycline stain¹⁵ (Fig. 3 inset b). Calcium partition inside the gels is likely to result from a Donnan equilibrium due to the polyanionic nature of their polymer matrix^{8,16–18}: at the normal pH of the sea water (pH 8.2), calcium attached to the gel matrix does not crystallize. But because the concentration of Ca^{2+} inside the gels must be higher and closer to critical saturation than the concentration of Ca^{2+} in sea water, slight increases of pH that do not produce precipitation in sea water can result in crystalline mineralization of the gel's matrix. Polyanion-mediated formation of mineral–polymer composites has been observed inside alginate microgels¹⁹, in the Golgi of coccolithophorid algae²⁰, and in other biomineralization processes^{18,21}. Hence, increased pH in sea water (8.5 to 9)—as found in microenvironments during high productivity²²—could induce crystalline mineralization of marine microgels. Crystalline mineralization could result in increased sedimentation rate, and may account for the formation of calcified aggregates and sediments (Fig. 4), such as mineral grains in marine aggregates, mineral-based marine snow, whittings and

oids^{6,23}. In hydrated microgels, polarization microscopy also reveals the presence of crystalline structures (data not shown). X-ray microanalysis (Fig. 3) indicates that this crystalline mineral–polymer composite is rich in calcium and low in phosphorus, suggesting the presence of calcium carbonate rather than calcium phosphate.

Images of fully hydrated gels were obtained by ESEM; (Fig. 3 inset a). This method complements the DLS and flow cytometry measurements, providing a precise, independent validation of the size of the gels. It also serves as a control for microbial contamination and for artefacts produced by other traditional electron microscopy methods that require chemical fixation. Histochemical staining with alcian blue, Ponceau S, Schiff stain, lectins, Nile red and fluorescently labelled antibodies against ribulose-1,5-biphosphate carboxylase/oxygenase (Rubisco)²⁴, indicate that the polymer matrix of these microgels contains carbohydrates, proteins (Rubisco) and lipids (data not shown).

A typical feature of all polymer gels is that they can undergo

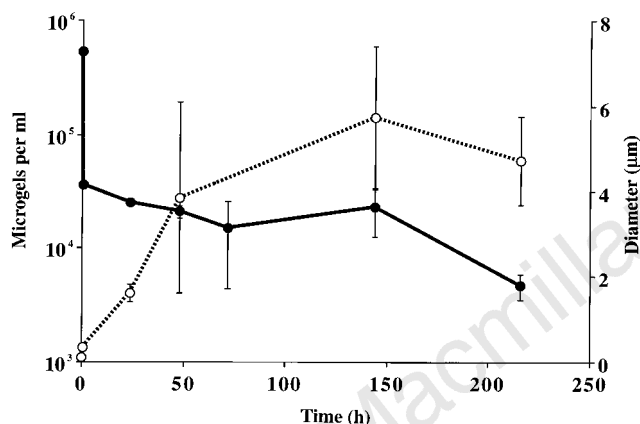


Figure 2 The concentration and size of marine microgels as a function of time of assembly. Data were obtained by flow cytometry in three replicates of 0.22-μm-filtered Puget Sound sea water. Although size increases (empty circles), the particle concentration in the sea water shows a corresponding decrease (filled circles). Because the volume of a particle is proportional to the cube of its diameter, the initial decrease of particle concentration should not be accompanied by a significant increase in particle size. The rate of decrease in microgel concentration is therefore expected to be faster than the rate of increase in microgel diameter. Each point corresponds to the average $\pm$ s.d. of two measurements on three replicate samples of sea water.

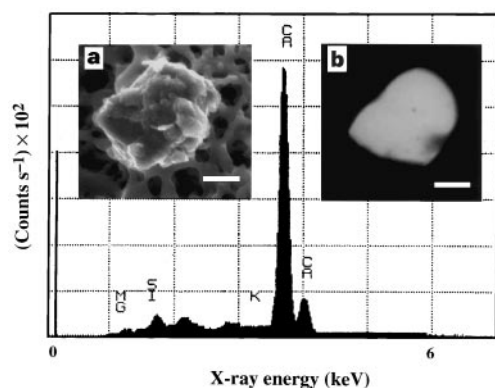


Figure 3 Averaged X-ray spectra showing the elemental composition inside a marine microgel. We note the large peak of calcium and the absence of phosphorus. Inset a shows an ESEM micrograph of a hydrated marine microgel (scale bar, 2 μm). Inset b reveals the presence of bound calcium in a microgel stained with chlortetracycline (scale bar, 2 μm).

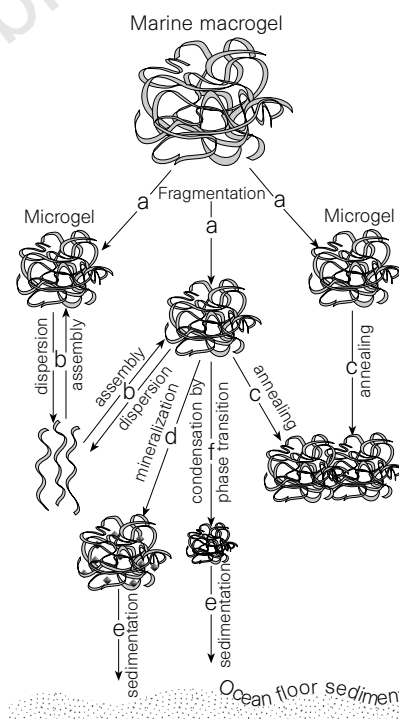


Figure 4 Processes transforming free marine polymers and polymer gels. Marine polymer microgels can result from partial fragmentation of gels secreted by phytoplankton or other organisms (a), or from assembly of free marine polymer (DOM) (b). The process of assembly is reversible, and microgels can be readily redispersed via physical, chemical or biological processes (b). Marine microgels can aggregate, and due to their tangled topology they can also 'anneal' to each other, forming larger stable polymer gels (c). Annealing is the process whereby polymers from one gel can diffuse and interpenetrate neighbouring gels, forming larger, interwoven networks. Because of their polyanionic properties, marine microgels concentrate calcium and can undergo crystalline mineralization (d), increasing their density and sedimentation rate (e). A less likely outcome in the ocean, but one that cannot be ruled out *a priori* is that microgels could undergo polymer gel phase transition by high pressure, low temperature and/or the presence of organic pollutants (f). The resulting high density of condensed microgels could also increase their vertical flux to the sediment (e). Both crystalline mineralization and/or condensation could potentially make microgels refractory to biodegradation.

reversible phase transitions from condensed to hydrated phases¹¹. By lowering the pH to 4.5 the microgels described here can also undergo a swelling/condensation transition from hydrated to condensed phase, as revealed by their volume change. The process is reversible and shows high apparent cooperativity, as shown by the steep sigmoidal swelling/condensation transitions depicted in Fig. 1b. The finding that condensation of marine polymer gels takes place at pH 4.5, the pK_a of carboxylic acid groups, is consistent with observations indicating that one out of every six carbon residues in DOM corresponds to a carboxylic group²⁵. These phase transition experiments were not designed to imitate conditions found in the ocean but to demonstrate the polymer gel nature of the networks assembled from DOM polymers. However, polymer gel phase transitions can be triggered by a variety of stimuli that may be relevant to the ocean, including high pressure, low temperature, or the presence of organic pollutants. The effect of these environmental conditions remains to be studied. When hydrogels undergo phase transition, their polymer network collapses into a dense, non-porous array that can exclude (or entrap) species of relative molecular mass as low as 300–500 (ref. 19). Biodegradable species can become refractory by forming, or by being entrapped in, microgels that exclude bacteria or enzymes via topological hindrance^{26,27}. Diffusional exclusion has been proposed to explain the unusual diagenetic stability of algaenans (algae-derived organic matter) and their high efficiency as carbon sinks²⁸. Condensation can also increase the density and sedimentation rate of these gels, providing an additional route for their removal from the water column (Fig. 4).

The present data show that without chemical reactions DOM polymers can entangle spontaneously to form POM microgels. According to Edward's theory (ref. 10), the probability of spontaneous assembly of tangled (non-covalently crosslinked) networks and the stability of these networks increase with the second power of the polymer length¹⁰. The dynamics of assembly of polymers in sea water, which contains longer chains than those only from the DOM pool, should therefore be faster, and the assembled gels should be more stable. Conversely, the breaking of marine polymers by ultraviolet photolysis²⁹ or by enzymatic degradation results in progressively shorter chains and correspondingly unstable gels. Polymer gel theory¹⁰ thus provides a simple explanation for the abundance and old age of low-molecular-weight DOM^{2–4}, whereas larger biopolymers are more likely to assemble to form stable POM gels, small polymers remain in the DOM pool owing to their low probability of stable assembly. As any kind of polymer shortening (physical, chemical or biological) decreases the chances of assembly into stable POM microgels, it will in general also reduce the accessibility to biodegradation. However, instead of increasing the bioavailability of organic matter, the formation of POM microgels could also have the opposite effect of preventing biodegradation: crystalline mineralization or phase-transition-induced condensation can turn POM microgels into refractory networks that may be more susceptible to removal from the water column through sedimentation (Fig. 4).

In the experimental conditions of this study, the assembly of DOM polymers results from collisions due to diffusional mobility. The influence on microgel assembly of physical (convective or turbulent flows, high pressures, or low temperatures), chemical or biological factors, or the presence of longer chain polymers (in this case excluded by the 0.22- μ m filtration), as found in the ocean, remains to be investigated. Although the detailed mechanisms giving rise to the transformation and degradation of marine organic matter remain controversial^{3,4}, the application of polymer gel physics promises to provide new answers. □

Methods

Water sampling and filtration. Seawater samples were collected in Puget Sound (WA, USA), the North Pacific and the Arctic. The samples were gravity

filtered through a GF/F fibreglass membrane and a 0.22- μ m Millipore tandem (pre-washed with 0.1 N HCl), transported to the laboratory in clean sealed bottles, and stored in the dark at 4 °C, for 1–6 weeks.

Particle sizing. The assembly of microgels was monitored by measuring particle size by both homodyne dynamics laser scattering and by flow cytometry. Seawater samples were shaken, and refiltered through a 0.22- μ m Millipore membrane (pre-washed with 0.1 N HCl). Aliquots (10-ml) were then poured directly into scattering cells. The scattering cells were positioned in the goniometer of a Brookhaven laser spectrometer (Brookhaven Instruments, NY, USA). Polymer assembly was monitored for 10 d, by analysing the scattering fluctuations detected at a 45° scattering angle. The autocorrelation function of the scattering intensity fluctuations was averaged over a 10-min sampling time, using a Brookhaven BI 9000AT autocorrelator. Particle size distribution was calculated by the CONTIN method³⁰. In parallel experiments, the polymer assembly was inhibited by chelating Ca^{2+} and Mg^{2+} by addition of 10 mM EDTA to the sea water. Control experiments were conducted in replicate samples containing 0.02% sodium azide to rule out microbial contamination. Measurements of microgel size were also performed in the same seawater samples, using a Coulter Epics Profile flow cytometer. In this case, the assembled gels were labelled with Nile red. Performance of these two particle sizing methods is complementary: whereas flow cytometry renders good estimates of large (>1 μ m) particle size, DLS yields reliable sizing in the submicrometre range. Calibrations of both methods were conducted using standard monodisperse suspensions of latex microspheres ranging from 50 nm to 6 μ m (Polysciences Inc., PA, USA).

Environmental scanning electron microscopy (ESEM). This powerful electron microscopy method enables materials—including living cells—to be studied at electron microscopy resolution (40 Å) while still fully hydrated. ESEM provides a non-destructive, non-invasive tool to investigate hydrated polymer networks in their native conformations. The present studies were conducted in assembled microgels retained on Millipore filters using an ElectroScan ESEM (model 2020, ElectroScan, MA, USA).

X-ray microanalysis. Microgels obtained by assembly of marine polymers from the DOM pool were fixed with a solution containing 0.5% glutaraldehyde, 0.25 M sucrose, and 0.1 M sodium cacodylate at pH 7.4, 20 °C. The samples were dehydrated in ethanol, and dried in fluorocarbon (Peltri II, Ted Pella, CA, USA). The specimens were sputter-coated with gold-palladium, and examined in a JEOL 840A (Japan) scanning electron microscope with energy dispersive spectroscopy capabilities.

Polymer gel phase transition. Swelling/condensation transitions were investigated at 20 °C by monitoring the volume changes of the microgels in sea water at pH ranging from 8 to 2.

Received 27 February; accepted 10 December 1997.

- Hedges, J. I. Global biogeochemical cycles: progress and problems. *Mar. Chem.* **39**, 67–93 (1992).
- Guo, L. & Santschi, P. H. Composition and cycling of colloids in marine environments. *Rev. Geophys.* **35**, 17–40 (1997).
- Kepkay, P. E. Particle aggregation and the biological reactivity of colloids. *Mar. Ecol. Prog. Ser.* **109**, 293–304 (1994).
- Amon, R. M. W. & Benner, R. Bacterial utilization of different size classes of dissolved organic matter. *Limnol. Oceanogr.* **41**, 41–51 (1996).
- Wells, M. L. & Goldberg, E. D. The distribution of colloids in the North Atlantic and Southern Oceans. *Limnol. Oceanogr.* **39**, 286–302 (1994).
- Cowen, J. P. & Holloway, C. F. Structural and chemical analysis of marine aggregates: in situ macrophotography and laser confocal and electron microscopy. *Mar. Biol.* **126**, 163–174 (1996).
- Allredge, A. L., Passow, U. & Logan, E. B. The abundance and significance of a class of large, transparent organic particles in the oceans. *Deep-Sea Res.* **40**, 1131–1140 (1993).
- Benedetti, M. F., Van Riemsdijk, W. H. & Koopal, L. K. Humic substances considered as a heterogeneous Donnan gel phase. *Environ. Sci. Technol.* **30**, 1805–1813 (1996).
- de Gennes, P. G. & Leger, L. Dynamics of entangled polymer chains. *Annu. Rev. Phys. Chem.* **33**, 49–61 (1982).
- Edwards, S. F. The theory of macromolecular networks. *Biorheology* **23**, 589–603 (1986).
- Li, Y. & Tanaka, T. Phase transitions of gels. *Annu. Rev. Mater. Sci.* **22**, 243–277 (1992).
- Stordal, M. C., Santschi, P. H. & Gill, G. A. Colloidal pumping: evidence for the coagulation process using natural colloids tagged with ²³⁰Hg. *Environ. Sci. Technol.* **30**, 3335–3340 (1996).
- Verdugo, P. Polymer gel phase transition in condensation-decondensation of secretory products. *Adv. Polym. Sci.* **110**, 145–156 (1994).
- Aluwihare, L. I., Repeta, D. J. & Chen, R. F. A major biopolymeric component to dissolved organic carbon in surface sea water. *Nature* **387**, 166–169 (1997).
- Caswell, A. H. & Hutchison, J. D. Visualization of membrane bound cations by a fluorescent technique. *Biochem. Biophys. Res. Commun.* **42**, 43–49 (1971).
- Katchalsky, A., Lifson, S. & Eisenberg, H. J. Equation of swelling for polyelectrolyte gels. *J. Polym. Sci.* **7**, 571–574 (1951).
- Tam, P. Y. & Verdugo, P. Control of mucus hydration as a Donnan equilibrium process. *Nature* **292**, 340–342 (1981).
- Crenshaw, M. A. in *Biological Mineralization and Demineralization* (ed. Nancollas, G. H.) 243–257 (Springer, Berlin, 1982).

19. Verdugo, P., Orellana, M. V. & Freitag, C. The secretory granule as a biomimetic model for drug delivery. *Proc. 22nd Int. Symp. on Controlled Release of Bioactive Materials* 22, 25 (1995).
20. Marsh, M. E. Polyanion-mediated mineralization-assembly and reorganization of acidic polysaccharides in the Golgi system of a cocolithophorid algal during mineral deposition. *Protoplasma* 177, 108–122 (1994).
21. Addadi, L., Moradian, J., Shay, E., Maraudas, N. G. & Weiner, S. A chemical model for the cooperation of sulfates and carboxylates in calcite crystal nucleation: relevance to biomineralization. *Proc. Natl. Acad. Sci. USA* 84, 2732–2736 (1987).
22. Lubbers, G. W., Gieskes, W. W. C., del Castillo, P., Salomons, W. & Bril, J. Manganese accumulation in the high pH microenvironment of *Phaeocystis* sp. (Haptophyceae) colonies from the North Sea. *Mar. Ecol. Prog. Ser.* 59, 285–293 (1990).
23. Morse, J. W. & Mackenzie, F. T. *Geochemistry of Sedimentary Carbonates* (Elsevier, Amsterdam, 1990).
24. Orellana, M. V. & Perry, M. J. An immunoprobe to measure Rubisco concentrations and maximal photosynthetic rates of individual phytoplankton cells. *Limnol. Oceanogr.* 37, 478–490 (1992).
25. Benner, R., Pakulski, J. D., McCarthy, M., Hedges, J. I. & Hatcher, P. G. Bulk chemical characteristics of dissolved organic matter in the ocean. *Science* 255, 1561–1564 (1992).
26. Mayer, L. M. Surface area control of organic carbon accumulation in continental shelf sediments. *Geochim. Cosmochim. Acta* 58, 1271–1284 (1994).
27. Nishikawa, T., Akiyoshi, K. & Sunamoto, J. Supramolecular assembly between nanoparticles of hydrophobized oligosaccharide and soluble protein complexation between the self-aggregate of cholesterol-bearing pullulan and α -chymotrypsin. *Macromolecules* 27, 7654–7659 (1994).
28. Largeau, C. in *Role of Nonliving Organic Matter in the Earth's Carbon Cycle* (eds Zepp, R. G. & Sonnet, C.) 275–292 (Wiley, New York, 1995).
29. Kieber, R. J., Zhou, X. & Mopper, K. Formation of carbonyl compounds from UV-induced photodegradation of humic substances in natural waters: fate of riverine carbon in the sea. *Limnol. Oceanogr.* 35, 1503–1515 (1990).
30. Provancher, S. W. A constrained regularization method for inverting data represented by linear algebraic or integral equations. *Comput. Phys. Commun.* 27, 213–227 (1982).

Acknowledgements. We thank M. J. Perry, P. Jumars and J. Hedges for their comments and suggestions. This work was supported by the US NSF (Polar Program) and the Royalty Research Fund (University of Washington).

Correspondence should be addressed to P.V.

Influence of oxygen exposure time on organic carbon preservation in continental margin sediments

Hilairy E. Hartnett, Richard G. Keil, John I. Hedges & Allan H. Devol

School of Oceanography, Box 357940, University of Washington, Seattle, Washington 98195-7940, USA

Today, over 90% of all organic carbon burial in the ocean occurs in continental margin sediments¹. This burial is intrinsically linked to the cycling of biogeochemically important elements (such as N, P, S, Fe and Mn) and, on geological timescales, largely controls the oxygen content of the atmosphere^{2–4}. Currently there is a volatile debate over which processes govern sedimentary organic carbon preservation^{5–8}. In spite of numerous studies demonstrating empirical relationships between organic carbon burial and such factors as primary productivity⁹, the flux of organic carbon through the water column¹⁰, sedimentation rate^{11,12}, organic carbon degradation rate¹³, and bottom-water oxygen concentration^{8,14}, the mechanisms directly controlling sedimentary organic carbon preservation remain unclear. Furthermore, as organic carbon burial is the process that, along with pyrite burial¹⁵, balances O₂ concentrations in the atmosphere, it is desirable that any mechanism proposed to control organic carbon preservation include a feedback buffering atmospheric oxygen concentrations over geological time. Here we compare analyses of sediments underlying two regions of the eastern North Pacific Ocean, one which has oxygen-depleted bottom waters and one with typical oxygen distributions. Organic carbon burial efficiency is strongly correlated with the length of time accumulating particles are exposed to molecular oxygen in sediment pore waters. Oxygen exposure time effectively incorporates other proposed environmental variables^{8–14}, and may exert a direct control on sedimentary organic carbon preservation and atmospheric oxygen concentrations.

Marine sediments and overlying waters were sampled during

three cruises off Washington State and two cruises off the north-western coast of Mexico. There is a prominent O₂-deficient zone over the Mexican margin between 150 and 600 m water depth (Fig. 1a) in which dissolved O₂ concentrations are often too low to detect by conventional techniques. In contrast, the Washington margin has a weaker O₂-minimum zone with concentrations decreasing to ~20 $\mu\text{mol l}^{-1}$ at 800 m. Both margins have a range in sediment accumulation rates (see Table 1). Ocean-colour satellite images from the Coastal Zone Color Scanner suggest that throughout the year, the Mexican margin study-site has lower pigment concentrations (and by inference, lower primary production rates) than the Washington coast site. The sharp contrasts between these two regions help to distinguish factors affecting sedimentary organic carbon (OC) preservation. For example, if primary production is the controlling factor, the Washington margin should have a higher degree of OC preservation. Alternatively, if sediment accumulation rate is the controlling factor, deposits with similar accumulation rates should have comparable degrees of OC preservation in both regions. Finally, if length of oxygen exposure dictates the extent of preservation, then the Mexican margin should have a greater degree of OC preservation.

Sediment samples were taken along transects that ran perpendicular to the shore from depths of ~100–1,000 m. Organic carbon contents in Washington margin sediments were <2.0 wt%, whereas those on the Mexican margin were as high as 12.0 wt% (Fig. 1b). This difference alone suggests the strongly oxygen-deficient zone off the coast of Mexico increases OC preservation. Mexican slope sediments, which had the lowest overlying water oxygen concentrations, had higher OC burial rates and lower rates of OC oxidation than did inshore shelf sediments (Table 1). Washington slope sediments had lower rates of both OC oxidation and burial than did Washington shelf sediments. Although OC burial rates for the Washington and Mexican margins were roughly similar, OC oxidation rates on the Washington margin were up to a factor of five higher than for the Mexican margin.

Organic-carbon burial efficiency has been used as an indicator of the extent of OC preservation in sediments^{10,12,16,17}. We define burial efficiency as the burial rate of OC below 15 cm, expressed as a

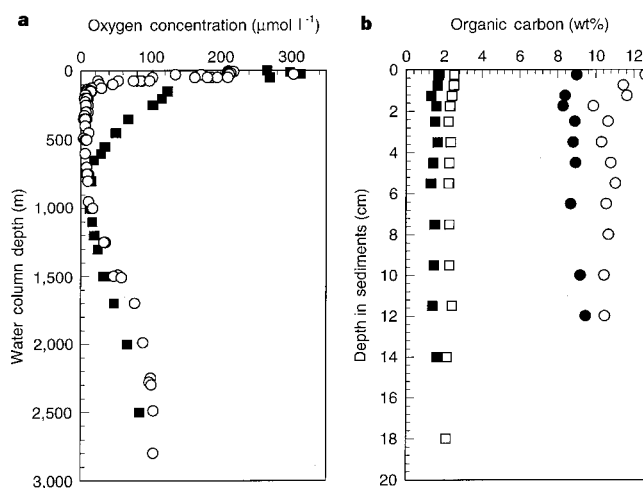


Figure 1 Water column O₂ and sediment carbon contents for study sites from the Washington and Mexican margins. **a**, Dissolved O₂ concentration as a function of water depth for the continental margins of Washington State (squares) and northwestern Mexico (circles). Note that the Mexican O₂ concentrations between 150 and 600 m depth are indistinguishable from zero. **b**, Weight per cent organic carbon as a function of depth in sediments for representative stations from the Washington (empty squares, 630 m; filled squares, 120 m) and Mexican (empty circles, 620 m; filled circles, 150 m).

percentage of the total OC input to the sea floor. Total OC input is the sum of OC burial and benthic OC loss due to oxidation. Organic carbon burial is given by $(\%OC)(S_{acc})$ where $\%OC$ is weight per cent organic carbon and S_{acc} is the sediment accumulation rate. $S_{acc} = S\rho(1 - \phi)$ where S is the linear sedimentation rate (cm yr^{-1})¹⁸, ρ is the dry sediment density (2.5 g cm^{-3}) and ϕ is the porosity. Carbon oxidation rates were calculated as the sum of the benthic O_2 consumption rate, the denitrification rate and the measured ^{35}S -labelled sulphate reduction rate¹⁹. Minimum and maximum estimates of the total carbon oxidation rate were determined, depending on the extent to which O_2 was assumed to re-oxidize the products of sulphate reduction; that is, no re-oxidation of sulphide by oxygen or the maximum possible sulphide oxidation. Uncertainty in the burial efficiency ($\sim 15\%$ as determined by a Monte Carlo error propagation technique²⁰) is due mainly to uncertainty in carbon oxidation rates and bulk sediment accumulation rates. Burial efficiency is quantitative and avoids treatment of OC preservation in terms that are mathematically interdependent (for example, sediment accumulation and OC accumulation) or mechanistically presumptive. Within our data set, sediments with higher bottom-water O_2 concentrations had lower burial efficiencies; the highest burial efficiencies ($>50\%$) occurred in regions with undetectable bottom-water oxygen concentrations (for example, Mexican slope). Other investigators have shown that where burial efficiencies are particularly low ($<2\%$), as is the case in the deep Pacific Ocean^{21,22}, bottom-water oxygen concentrations are generally high: this again suggests that through some mechanism O_2 is affecting the extent to which OC is permanently buried. Oxygen concentration alone, however, is an instantaneous, static measurement and does not reflect the dynamic processes controlling OC preservation.

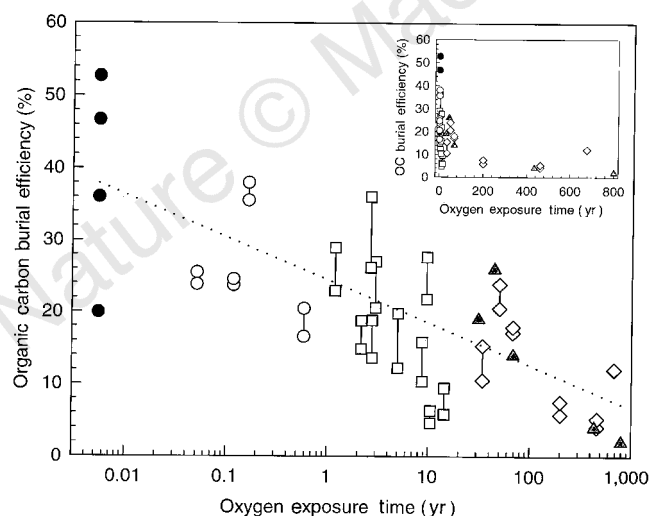


Figure 2 Organic carbon burial efficiency as a function of oxygen exposure time. Symbols are as follows: the Mexican shelf, 100–150 m (empty circles); Mexican slope oxygen-deficient zone, 150–1,000 m (filled circles); the Washington shelf and upper slope, 100–600 m (shaded squares); Washington lower slope (J.I.H. *et al.*, manuscript in preparation) 600–2,500 m (empty diamonds); and the California margin¹⁷, 1,500–3,500 m (shaded triangles). O_2 exposure time is plotted on a log-linear scale so that variations in burial efficiency at short exposure times can be distinguished; the dotted line is a least-squares fit to the data. For Mexican slope stations with zero O_2 exposure time, the data are plotted at 0.003 yr (<2 d) as the log scale cannot accommodate a value of zero. Where we calculate a maximum and minimum total carbon oxidation rate, we present the corresponding range in carbon burial efficiency as two connected symbols. Inset, as main figure but with a linear-linear scale to demonstrate the exponential nature of the relationship.

In contrast, the total contact time of O_2 with depositing organic matter should integrate any O_2 effect, and can be estimated for most depositional environments. We determined O_2 exposure time as the depth of oxygen penetration divided by the linear sedimentation rate. These exposure-time calculations have an uncertainty of $\pm 50\%$, due mainly to the uncertainty in the oxygen penetration depth. Washington margin O_2 penetration depths range from 2 to 10 mm. On the Mexican shelf, O_2 penetration is much shallower (~ 1 mm). The Washington margin O_2 penetration depth was determined directly from microelectrode profiles of porewater O_2 (ref. 23). For the four Mexican shelf stations, O_2 penetration depth was calculated from the benthic O_2 flux¹⁹; that is, O_2 penetration depth is given by $[(D_{\text{O}_2}^s)(\text{O}_{2\text{BW}})]/(\text{benthic } \text{O}_2 \text{ flux})$, where $\text{O}_{2\text{BW}}$ is the bottom water O_2 concentration, $D_{\text{O}_2}^s \equiv (D_{\text{O}_2}^m)/(\phi^2)$, $D_{\text{O}_2}^s$ is the sedimentary diffusion coefficient, and $D_{\text{O}_2}^m$ is the molecular diffusion coefficient for O_2 at the *in situ* temperature²⁴. This parametrization assumes a simple linear O_2 gradient, and yields extremely shallow (<2 mm) Mexican margin O_2 penetration depths. The alternative assumption of an exponential O_2 gradient does not change the penetration depth by more than 50%. Mexican slope sediments had no measurable oxygen flux and thus exposure times of zero. The activity of benthic macrofauna may also influence O_2 exposure time as porewater irrigation can increase O_2 fluxes to several times those calculated from the gradient-driven diffusive flux²³. Bioturbation does not have a large effect on the exposure time as defined here because random physical mixing moves sediments into and out of the oxic zone at the same rate, so the average exposure time for any one particle will be the same with or without mixing. This is a minimum estimate of exposure time that does not reflect any past contact with O_2 at a previous location during winnowing or horizontal transport in the benthic boundary layer. In determining exposure times, oxygen is the most relevant oxidant in the sequence of electron acceptors for most continental margin sediments. In both regions, exposure to porewater nitrate is similar, and denitrification rates are small relative to O_2 consumption rates. Exposure to sulphate is also similar in both regions, where dissolved sulphate penetrates hundreds of centimetres into the sediment column (SO_4^{2-} exposure times of the order of thousands of years), and therefore cannot account for the observed difference in OC preservation on the Washington and Mexican margins.

Comparison of the two northeast Pacific margins indicates that the highest OC burial efficiencies occur in those regions with the shortest O_2 exposure times (Fig. 2). We have calculated exposure times for the Mexican and Washington margins which range from approximately zero to 20 years (Table 1). To examine longer O_2 exposures, we used similarly derived data for OC burial, O_2 penetration and sedimentation rate to compare our results to those for sediments from the deep California margin¹⁷ and from the lower Washington slope (J.I.H. *et al.*, manuscript in preparation). These data extend the range of calculated exposure times to >800 years (Fig. 2). There is a strong correlation ($r^2 > 0.62$) between the OC burial efficiency and the logarithm of O_2 exposure over the entire range of times. This range in exposure times includes sediments that are overlain by anoxic water columns (Mexican slope) as well as typical oxic deep-sea sediments (deep Washington slope). These data suggest a continuum extending from regions with short exposure time and a high degree of preservation to regions with long oxygen exposure times and less preservation. In general, sediments with longer exposure times have lower OC contents (Fig. 1b), as well as lower OC burial efficiencies (Fig. 2).

Oxygen exposure time more generally relates to OC burial than previously suggested individual variables because the other proposed factors (that is, primary production rate, rain rate of OC, overall sedimentation rate, and bottom-water O_2 content) will each influence the oxygen exposure time of accumulating organic matter. Increases in either the primary production rate or the OC rain rate will increase sediment oxygen demand. The consequent steepening

Table 1 Measured and calculated parameters for Washington and Mexican margin sediments

	Bottom water O ₂ ($\mu\text{mol l}^{-1}$)	Organic carbon (wt%)	Sediment accumulation rate* ($\text{mg cm}^{-2} \text{yr}^{-1}$)	Organic carbon burial ($\mu\text{mol cm}^{-2} \text{yr}^{-1}$)	Organic carbon oxidation ($\mu\text{mol cm}^{-2} \text{yr}^{-1}$)	Burial efficiency (%)	O ₂ exposure time (yr)
Washington Shelf†							
Average (n = 8)	92.8	1.31	102	117	675	15.1	3.92
Range	(77–106)	(0.55–1.8)	(61–130)	(28–169)	(506–790)	(3.75–17.8)	(1.1–10.5)
Upper slope							
Average (n = 6)	71.5	1.59	66.9	90.1	316	25.7	6.42
Range	(38–104)	(0.5–2.8)	(37–100)	(27–210)	(91–607)	(4.69–42.7)	(1.4–14.4)
Mexico Shelf‡							
Average (n = 4)	15.4	4.7	14.9	50.9	183	23.3	0.252
Range	(3–0)	(2.9–7.1)	(9.1–25.6)	(195–392)	(157–200)	(18.8–26.3)	(0.051–0.587)
Slope							
Average (n = 4)	5.3	10.1	6.87	61.7	91.3	38.2	0.032
Range	(0.0–12)	(7.5–12.8)	(4.1–12.7)	(160–845)	(55–121)	(19.9–53.1)	(0.0–0.16)

* Sediment accumulation rates and wt% OC determinations were made on the same cores; these cores were collected at the stations where benthic flux chamber measurements were made.

† Washington shelf and upper-slope stations had depths in the range 0–200 m and 200–600 m, respectively.

‡ The Mexican shelf extends to ~150 m and the Mexican slope stations range from 150 to 1,000 m.

of the O₂ gradient will lead to a shallower O₂ penetration depth and decreased O₂ exposure time. Likewise, variations in the mineral sedimentation rate will affect the rate at which material moves through the oxic horizon at the sediment surface, and thus alter O₂ exposure time. Finally, if all other factors remain constant, changes in the bottom-water O₂ concentration will result in corresponding changes in O₂ exposure time. The correlation often observed between these other parameters and OC preservation^{10,12,25} probably reflects their individual influence on oxygen exposure time rather than a direct influence on OC burial. The heart of the preservation controversy lies in the fact that no single factor seems to control preservation under all conditions. The O₂ exposure time hypothesis assumes that a certain amount of OC degradation in sediments requires oxygen. Although not mechanistically explicit, the hypothesis is consistent with molecular-level mechanisms that have been proposed requiring the presence of O₂ to degrade carbon-rich, hydrolysis-resistant, organic matter^{1,26}. Our use of a mathematically independent measure of preservation (burial efficiency) and a multivariate controlling parameter (oxygen exposure time) allows a realistic comparison of the extent of preservation under a variety of conditions, and in different regions, while accounting for the combined influence of several factors contributing to OC preservation.

The control of OC preservation by length of exposure to O₂ provides a critical negative feedback between the burial of reduced materials in sediments and the O₂ content of the oceans and atmosphere. For example, large increases in organic matter burial over time would increase atmospheric O₂ concentrations, leading (via atmospheric exchange) to higher O₂ concentrations in ocean bottom water. Oxygen fluxes across the sediment–water interface would increase proportionately, leading to deeper O₂ penetration and, therefore, longer O₂ exposure times and more OC remineralization. Although it has long been suggested that an O₂-based feedback must exist⁴, this is the first clear evidence that OC burial can be linked directly to a control on atmospheric O₂ rather than indirectly via subaerial kerogen weathering on land. The exact microbiological and chemical mechanisms by which OC degradation is mediated in oxic versus anoxic sedimentary environments, however, remain to be elucidated. More extensive estimation of oxygen exposure, either through coupled rate and burial estimates (as described here) or through chemical changes in organic matter during diagenesis^{26,27}, should allow a better estimation of the extent to which oxygen exposure affects organic carbon burial worldwide. □

Received 21 October 1996; accepted 28 October 1997.

- Hedges, J. I. & Keil, R. G. Sedimentary organic matter preservation: an assessment and speculative synthesis. *Mar. Chem.* **49**, 81–115 (1995).

- Berner, R. A. Burial of organic carbon and pyrite sulfur in the modern ocean: Its geochemical and environmental significance. *Am. J. Sci.* **282**, 451–473 (1982).
- Hedges, J. I. Global biogeochemical cycles: progress and problems. *Mar. Chem.* **39**, 67–93 (1992).
- Holland, H. D. *The Chemistry of the Atmosphere and Oceans* (Wiley, New York, 1978).
- Canfield, D. E. Sulfate reduction and oxic respiration in marine sediments: implications for organic carbon preservation in euxinic environments. *Deep-Sea Res.* **36**, 121–138 (1989).
- Henrichs, S. M. Early diagenesis of organic matter in marine sediments: progress and perplexity. *Mar. Chem.* **39**, 119–149 (1992).
- Pedersen, T. F., Shimmield, G. B. & Price, N. B. Lack of enhanced preservation of organic matter in sediments under the oxygen minimum on the Oman margin. *Geochim. Cosmochim. Acta* **56**, 545–551 (1992).
- Paropkari, A. L., Prakash Babu, C. & Mascarenhas, A. A critical evaluation of depositional parameters controlling the variability of organic carbon in Arabian Sea sediments. *Mar. Geol.* **107**, 213–226 (1992).
- Calvert, S. E. & Pedersen, T. F. In *Productivity, Accumulation and Preservation of Organic Matter in Recent and Ancient Sediments* (eds Whelan, J. K. & Farrington, J. W.) 231–263 (Columbia Univ. Press, New York, 1992).
- Jahnke, R. A. Early diagnosis and recycling of biogenic debris at the seafloor, Santa Monica Basin, California. *J. Mar. Res.* **48**, 413–436 (1990).
- Müller, P. J. & Suess, E. Productivity, sedimentation rate, and sedimentary organic matter in the oceans—I. Organic carbon preservation. *Deep-Sea Res.* **A 26**, 1347–1362 (1979).
- Betts, J. N. & Holland, H. D. The oxygen content of ocean bottom waters, the burial efficiency of organic carbon, and the regulation of atmospheric oxygen. *Paleogeogr. Paleoclimatol. Paleocool.* **97**, 5–18 (1991).
- Emerson, S. In *The Carbon Cycle and Atmospheric CO₂: Natural Variations Archaeal to Present* (eds Sundquist, T. & Broecker, W. S.) 78–87 (Am. Geophys. Union, Washington DC, 1985).
- Demailson, G. J. & Moore, G. T. Anoxic environments and oil source bed genesis. *Org. Geochem.* **2**, 9–31 (1980).
- Berner, R. A. & Canfield, D. E. A new model for atmospheric oxygen over phanerozoic time. *Am. J. Sci.* **289**, 333–361 (1989).
- Henrichs, S. M. & Reeburgh, W. S. Anaerobic mineralization of marine sediment organic matter: Rates and the role of anaerobic processes in the oceanic carbon economy. *Geomicrobiol. J.* **5**, 191–237 (1987).
- Reimers, C. E., Jahnke, R. A. & McCorkle, D. C. Carbon fluxes and burial rates over the continental slope and rise off central California with implications for the global carbon cycle. *Glob. Biogeochem. Cycles* **6**, 199–224 (1992).
- Carpenter, R., Bennet, J. T. & Peterson, M. L. ²¹⁰Pb activities in and fluxes to sediments of the Washington continental slope and shelf. *Geochim. Cosmochim. Acta* **45**, 1155–1172 (1981).
- Devol, A. H. & Christensen, J. P. Benthic fluxes and nitrogen cycling in sediments of the continental margin of the eastern North Pacific. *J. Mar. Res.* **51**, 345–372 (1993).
- Bevington, P. R. & Robinson, D. K. *Data Reduction and Error Analysis for the Physical Sciences* (McGraw-Hill, New York, 1992).
- Jahnke, R. A., Emerson, S. R. & Murray, J. W. A model of oxygen reduction, denitrification and organic matter remineralization in marine sediments. *Limnol. Oceanogr.* **27**, 610–623 (1982).
- Martin, W. R., Bender, M., Leinen, M. & Orchard, J. Benthic organic carbon degradation and biogenic silica dissolution in the central equatorial Pacific. *Deep-Sea Res.* **38**, 1481–1516 (1991).
- Archer, D. & Devol, A. H. Benthic oxygen fluxes on the Washington shelf and slope: A comparison of *in situ* microelectrode and chamber flux measurements. *Limnol. Oceanogr.* **37**, 614–629 (1992).
- Lehrman, A. *Geochemical Processes: Water and Sediment Environments* (Wiley & Sons, New York, 1979).
- Calvert, S. E., Bustin, R. M. & Pedersen, T. F. Lack of evidence for enhanced preservation of sedimentary organic matter in the oxygen minimum of the Gulf of California. *Geology* **20**, 757–760 (1992).
- Keil, R. G., Hu, F. S., Tsamakis, E. C. & Hedges, J. I. Pollen in marine sediments as an indicator of oxidation of organic matter. *Nature* **369**, 639–641 (1994).
- Cowie, G. L. & Hedges, J. I. Biochemical indicators of diagenetic alteration in natural organic matter mixtures. *Nature* **369**, 304–307 (1994).

Acknowledgements. We thank B. Hales and J. Brandes for discussions, and D. Lambourn, E. Tsamakis, C. Thimsen, B. Paul, L. Balistrieri and numerous participants in the U.W. REU program for analytical assistance. This work was supported by NSF grants to A.H.D. and R.G.K.

Correspondence and requests for materials should be addressed to H.E.H. (e-mail: hartnett@ocean.washington.edu).

Oceanic transport of subpolar climate signals to mid-depth subtropical waters

Ruth G. Curry, Michael S. McCartney & Terrence M. Joyce

Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA

The spatial distributions of certain sea-surface properties, such as temperature, fluctuate on timescales from months to decades and in synchrony with the main regional atmospheric patterns comprising the global climate system¹. Although it has long been assumed that the ocean is submissive to the dictates of the atmosphere, recent studies raise the possibility of an assertive, not merely passive, oceanic role in which water-mass circulation controls the timescales of climate fluctuations^{2–6}. Previously held notions of the immutability of the physical and chemical characteristics of deep water masses are changing as longer time series of ocean measurements indicate that the signatures of varying sea-surface conditions are translated to deep waters^{4,7}. Here we use such time-series measurements to track signals 'imprinted' at the sea surface in the North Atlantic Ocean's subpolar Labrador Basin into the deep water of the subtropical basins near Bermuda, and infer an approximately 6-year transit time. We establish a geographic and temporal context for a portion of the long-term warming trend reported for mid-depth subtropical waters over the past 40 or so years^{8,9}, and we predict that waters at these depths will continue to cool well into the next decade.

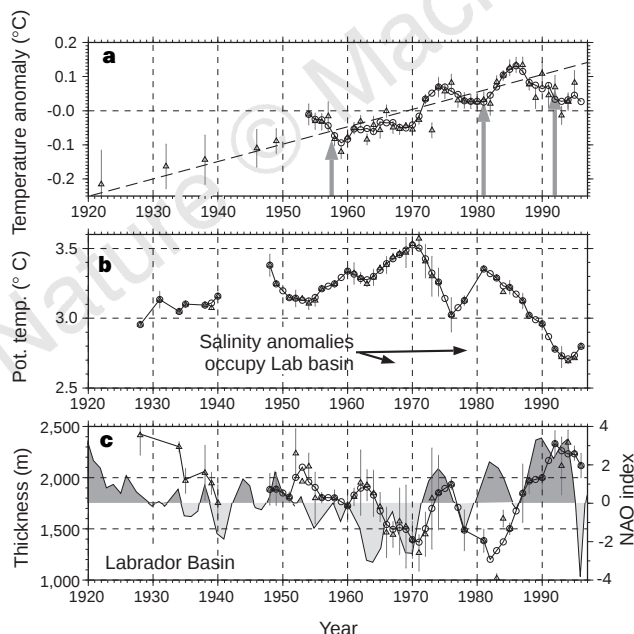


Figure 1 Evolution of subtropical and subpolar water masses through time. **a**, Time series of subtropical temperature anomaly in 1,500–2,500-m layer near Bermuda⁹. Triangles are annual mean, vertical bars are s.d. Circles and curve depict 3-year smoothed values. The broken line is a warming trend inferred from data⁹. Arrows indicate years from which section data were compared to deduce a similar warming trend^{8,10}. **b**, Potential temperature of LSW core (at 1,500 m) in its subpolar formation area. Means and standard deviation are as in **a**. Cross-hatched years delimit periods during which fresh water capped the Labrador basin. Data source is HydroBase²⁰. **c**, Thickness of LSW density layer ($\sigma_{1,500} = 34.62\text{--}34.72$) in its formation area. The NAO index of Hurrell¹⁴ is overplotted with high index years shaded dark and low index years shaded lightly.

Warming of the subtropical mid-depths has been deduced by a variety of techniques including the comparison of zonal^{8,10} and meridional⁹ temperature sections from different years, temperature-mapped on depth surfaces¹¹ from different time windows, and from analyses of the temperature–depth time series maintained near Bermuda^{9,12}. A trend analysis (broken line in Fig. 1a) of this time series, reproduced in Fig. 1a, implies a warming of 0.5 °C per century over the 70-year record⁹. The densely sampled years after 1954 show decadal oscillations superimposed on the long-term trend^{9,12}, and, by linking the time series analysis with the changes observed in temperature sections from different years^{8–10}, a 1,000-km scale variability at 10-year timescales is inferred for the western basin^{8,9}.

Over the same period, subpolar mid-depth conditions in the Labrador basin show a long-term warming trend from the 1920s to 1971 that ends abruptly in 1972 and is followed by a cooling trend that persists to the present¹³. A time series of temperature of the Labrador Sea Water (LSW) core measured in its source region (Fig. 1b) illustrates these trends. Formed by deep wintertime convection in the Labrador basin, the LSW is characteristically cold, fresh and very thick compared with other North Atlantic water masses of similar density. The time history of LSW thickness—defined here by the vertical distance between two density surfaces—approximately mirrors the temperature trends (Fig. 1c): the long period of warming before 1972 corresponds to a thinning of the layer, whereas the cooling years of the 1970s and after 1983 are accompanied by layer thickening. The thickness of the LSW layer is directly related to the intensity of wintertime convection: strong convection produces a thick layer, and weak convection results in a relatively thin layer.

The depth of Labrador basin convection is principally determined by the strength of the westerly winds and the occasional passage of freshwater surface anomalies^{4,7,13}. Also possibly involved are anomalies of SST and heat content carried from the subtropics into the subpolar gyre^{5,19}. An increase in wind strength removes more heat from the surface waters and deepens the extent of convection. This also results in a cooler overall LSW, because the increased heat loss at the sea surface is distributed downward as the water column overturns. The North Atlantic Oscillation (NAO) index¹⁴, which represents the relative strength of the westerlies, is overplotted on the LSW thickness axis (Fig. 1c) and shows trends similar to the LSW properties: declining NAO index (weakening winds), thinning and warming LSW from 1950s to 1970, a pulse of strong westerly winds (high index), LSW thickening and cooling in the early 1970s followed by weak westerlies (low index) and thin conditions in the late 1970s, then extremely strong westerlies, thick and cold LSW in the 1990s. There is a suggestion of perhaps a 2–4-year time lag between the atmospheric NAO and LSW thickness indices, which is indicative of the ocean's slower response time. The wind strength briefly loses its influence on LSW conditions during occasional salinity events: the thin LSW layers of 1968–72 and the early 1980s

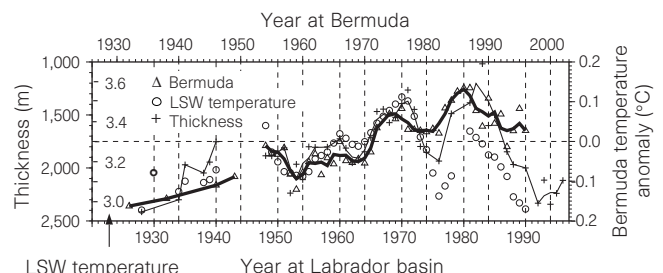


Figure 2 Bermuda temperature anomaly from Fig. 1a lagged by 6 years (triangles, thick curve) with LSW source thickness (thin curve and + symbols; axis has been inverted compared with Fig. 1c) and LSW core temperature (circles). After 1990, LSW temperatures are colder than 2.9 °C and are not visible in this plot.

correspond to buildups of extremely low surface salinities, which completely inhibited convective overturning^{4,7,13}. The first of these two events, referred to as the 'great salinity anomaly'¹⁵, occupied the Labrador basin during a time of low NAO index when weak westerlies would have resulted in weak convection anyway. Note that the warming and thinning trend spans two decades, mirroring the general decline of the NAO index of the 1950s and 1960s, whereas the low-salinity event occurs in 1968–72, only at the very end of the trend. However, the second event, the 'lesser great salinity anomaly'¹⁶, shut down convection in the early 1980s causing the LSW to become thinner despite strong westerlies associated with a relatively high NAO index. Both events ended with a strong increase in convection and a downward mixing of the chilled freshwater cap, which markedly lowered the LSW core temperature (Fig. 1b) and salinity¹².

The subpolar LSW enters the subtropics via two principal pathways: the deep western boundary current (DWBC) advects LSW southward to the tropics, whereas the Gulf Stream and North Atlantic Current transport it from the western boundary out into the basin interior¹⁷. The LSW in those flows is strongly stirred and mixed by the action of recirculations and eddies along these pathways. The subtropical basin-scale deep-water properties thus repre-

sent an advective–diffusive blending of the cold, fresh, thick LSW with other influences such as the warm, salty, thin Mediterranean Outflow Water (MOW); the resulting water mass, Upper North Atlantic Deep Water (UNADW), exhibits a temperature history that we can now relate to variations in the LSW source as follows. Over the past century, the volume of LSW flowing out of the subpolar gyre has increased/decreased as this water mass has thickened/thinned. A decrease in LSW entering the subtropics causes a relative ascendancy of the MOW characteristics and results in a warmer, saltier and thinner UNADW layer. Conversely, an increased volume of LSW results in fresher, cooler, thicker UNADW. The time required to advect and mix the LSW into the subtropics introduces a time delay to the link between these signals.

When a time lag is applied to the Bermuda temperature anomaly signal (from Fig. 1a), the subpolar and subtropical records are strikingly similar both over the long term (70 years) and at decadal timescales (Fig. 2). The coldest, but very sparse, Bermuda data before 1950 correspond to thick and cold LSW conditions. From 1922 to 1987 Bermuda exhibits a general warming, whereas the LSW exhibits an overall thinning from 1928 to 1983. The LSW began to thicken and cool around 1982, reaching extreme values in the 1990s; Bermuda temperatures began to cool in 1988. The subpolar LSW pulsed cooler and thicker for 4 years in the 1970s, which is echoed by a limited cooling at Bermuda from 1977 to 1982, but did not reverse the subtropical warming trend. Negative correlation between the subtropical temperature and subpolar thickness signals is strongest ($r < -0.6$) at lags of 5–7 years and implies that when subpolar convection is strong—and the LSW layer is thick—the subtropics follow about 6 years later with cooler UNADW temperatures. Conversely, weak convection and a thin LSW layer are followed by warmer UNADW. The lag is not precise because the basin-scale advective–diffusive mixing is not an exact process, and because earlier measurements in both time series were less abundant and also made with cruder technology.

To place this relationship into a geographic context, Fig. 3a maps the thickness of the LSW density layer ($\sigma_{1,500} = 34.62$ – 34.72) and Fig. 3b maps the temperature on a density surface ($\sigma_{1,500} = 34.7$) in the middle of that layer. We use density as opposed to depth surfaces because adjacent water masses mix predominantly along these isopycnal surfaces, which vary greatly in depth from the subpolar (500 m) to subtropical gyre (1,800 m). In Fig. 3a the red colours indicate thicknesses exceeding 2,000 m in the LSW source region. The green tongue (thickness < 500 m) stretching westwards across the subtropics delineates where MOW influences are strong; deep blue colours represent waters whose properties (thickness, temperature, salinity, and so on) are intermediate between the LSW and MOW sources. In Fig. 3b the LSW is identified by coldest temperatures (blue colours), the MOW shows as the red–yellow tongue, and the product of their mixing is green.

The remaining panels of Fig. 3 demonstrate the time-lagged impact of LSW temperature and thickness anomalies on the subtropics. Here we subdivided the post-1956 data into five periods, each spanning 7–8 years, and mapped the thickness and temperature differences in one time frame compared with its preceding time frame. The alternating yellow–red (thinning, warming) and blue–green (thickening, cooling) patches in the Labrador basin depict the waning and waxing through time of LSW in its source region. These patterns also show large areas where thickness changes are correlated with temperature changes—that is, where the layer thins temperatures are warming, and where the layer thickens temperatures are cooling—and depict where LSW exerts a strong influence. In the eastern basin the main impact of the LSW thickness anomalies are restricted to the area north of 30° N, whereas in the western basin they are intensified towards the west but extend down to the tropics. These patterns also show consecutive instances where temperature and thickness anomalies of one sign first appear in the Labrador basin, are quickly advected southwards and eastwards by

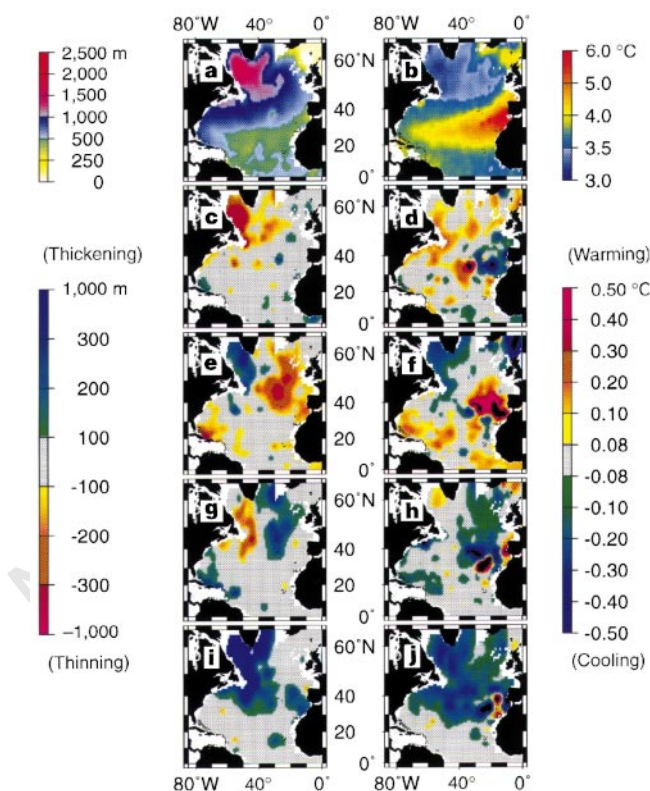


Figure 3 Anomalies of thickness and temperature propagate away from their subpolar source. **a**, Mean thickness of the density layer corresponding to Labrador Sea Water ($\sigma_{1,500} = 34.62$ – 34.72) over the entire North Atlantic. Data source is HydroBase²⁰. **b**, Potential temperature (climatological mean) on a density surface in the middle of the LSW layer ($\sigma_{1,500} = 34.67$). **c**, Change in thickness of LSW layer mapped in **a** between two time periods: 1966–72 compared with 1966–65. Blue–green colours correspond to layer thickening in the later time compared with the earlier; yellow–red colours indicate layer thinning. **d**, Change in temperature on the density surface mapped in **b** for the time periods compared in **c**. Blue–green colours correspond to cooling in the later time compared with the earlier; yellow–red colours indicate warming. **e**, **f**, Thickness and temperature change: 1973–79 compared with 1966–72. **g**, **h**, Thickness and temperature change: 1980–86 compared with 1973–79. **i**, **j**, Thickness and temperature change: 1987–94 compared with 1980–86.

the DWBC and North Atlantic Current (e.g. in Fig. 3c and d, where anomalies have same colour as in the Labrador basin) and then, one time frame later, anomalies of the same sign appear in both the western and eastern subtropical basins. This is consistent with the time lag implied by the Bermuda temperature/Labrador basin thickness correlation. The eastern subpolar gyre shows a similar delay, which is consistent with inferred influence times since the 1980s^{14,18}.

The time-delayed subtropical temperature patterns seem to represent the slow adjustment of the subtropical deep water's advective–diffusive balance to the waxing and waning LSW. As the LSW source thickens, its role in that balance strengthens and this is manifested as an eastwards and southwards erosion of the MOW influence on the subtropical deep water and thus cooler and fresher UNADW. The time required for the advective elements to circulate and mix the LSW into the subtropical basins results in a delayed appearance of the response. When the LSW is thinner than normal, the MOW is able to exert more influence on the advective–diffusive balance, and this appears as a westwards and northwards extension of its warm, salty and thin characteristics.

Understanding the nature of the subtropical temperature response and knowing that subpolar convection has been extreme from 1988 to 1995 prompts us to speculate that the subtropical mid-depths will continue to cool throughout the 1990s and the early 2000s. The cold, fresh and thick signal now invading the subtropics is quite pronounced in the bottom panels of Fig. 3 and will undoubtedly have profound effects on the UNADW properties. The colder, fresher UNADW arrived at Bermuda in August 1996. And its arrival in the DWBC at 26° N was noted by increased chlorofluorocarbon concentrations in 1996 by Miami researchers¹⁹ who have been monitoring this region for about a decade. If the NAO index were to remain persistently high and subpolar convection strong, then the 40-year warming of the subtropical mid-depths would reverse to a substantial cooling. A switch from the high NAO of the 1980s and early 1990s to a persistent low index with weak subpolar convection, however, would again reduce the amount of LSW entering the subtropics and this impending cooling could appear as just another decadal oscillation in the long-term record. Such a switch might be in progress, for the NAO index has been low for two consecutive winters: 1995/96 and 1996/97. □

Received 17 March; accepted 14 October 1997.

- Deser, C. & Blackmon, M. Surface climate variations over the North Atlantic Ocean during winter: 1900–1989. *J. Clim.* **6**, 1743–1753 (1993).
- Kushnir, Y. & Held, I. M. Equilibrium atmospheric response to North Atlantic SST anomalies. *J. Clim.* **9**, 1208–1220 (1996).
- Griffies, S. M. & Bryan, K. A predictability study of simulated North Atlantic multidecadal variability. *Clim. Dyn.* **13**(7/8), 459–487 (1997).
- Dickson, B. From the Labrador Sea to global change. *Nature* **386**, 649–650 (1997).
- Sutton, R. T. & Allen, M. R. Decadal predictability of North Atlantic Sea surface temperature and climate. *Nature* **388**, 563–565 (1997).
- McCartney, M. S. Is the ocean at the helm? *Nature* **388**, 521–522 (1997).
- Dickson, R., Lazier, J., Meincke, J., Rhines, P. & Swift, J. Long-term coordinated changes in the convective activity of the North Atlantic. *Prog. Oceanogr.* **38**, 241–295 (1996).
- Bryden, H. L. *et al.* Decadal changes in water mass characteristics at 24° N in the subtropical North Atlantic Ocean. *J. Clim.* **9**, 3162–3186 (1996).
- Joyce, T. M. & Robbins, P. The long-term hydrographic record at Bermuda. *J. Clim.* **9**, 3121–3131 (1996).
- Roemmich, D. & Wunsch, C. Apparent changes in the climatic state of the deep North Atlantic Ocean. *Nature* **307**, 447–450 (1984).
- Levitus, S. Interdecadal variability of temperature and salinity at intermediate depths of the North Atlantic Ocean, 1970–1974 versus 1955–1959. *J. Geophys. Res.* **94**, 6091–6131 (1989).
- Houghton, R. W. Subsurface quasi-decadal fluctuations in the North Atlantic. *J. Clim.* **9**, 1363–1373 (1996).
- Lazier, J. R. in *Natural Climate Variability on Decade-to-Century Time Scales of Natural Climate Variability* (ed. Martinson, D. G. *et al.*) 295–302 (National Academy Press, Washington, DC, 1995).
- Hurrell, J. W. Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation. *Science* **269**, 676–679 (1995).
- Dickson, R., Meincke, J., Malmberg, S.-A. & Lee, A. The “Great Salinity Anomaly” in the northern North Atlantic 1968–1982. *Prog. Oceanogr.* **20**, 103–151 (1988).
- Belkin, I. M., Levitus, S. & Antonov, J. “Great Salinity Anomalies” in the North Atlantic. *Prog. Oceanogr.* (in the press).
- Talley, L. D. & McCartney, M. S. Distribution and circulation of Labrador Sea Water. *J. Phys. Oceanogr.* **12**, 1189–1205 (1982).
- Sy, A. *et al.* Surprisingly rapid spreading of newly formed intermediate waters across the North Atlantic Ocean. *Nature* **386**, 675–679 (1997).

- Molinari, R. *et al.* Fast-track for recently formed Labrador Sea Water: the Deep Western Boundary Current of the North Atlantic Ocean. *Deep-Sea Res.* (submitted).
- Curry, R. HydroBase: a database and tools for climatologic analysis. (Tech. Rep. WHOI-96-01, Woods Hole Oceanographic Inst, 1996).

Acknowledgements. This work was supported by the Atlantic Climate Change Program of NOAA's Climate and Global Change Division and by the National Science Foundation.

Correspondence and requests for materials should be addressed to R.G.C. (e-mail: rcurry@whoi.edu). Data available from flotsam.whoi.edu (128.128.28.187). cd pub/nature/curry_et_al.

The generation of plankton patchiness by turbulent stirring

Edward R. Abraham

National Institute of Water and Atmospheric Research, P.O. Box 14-901, Kilbirnie, Wellington, New Zealand

Diffusive processes are often used to represent the formation of spatial patterns in biological systems¹. Here I show how patchiness may be generated in planktonic ecosystems through non-diffusive advection. Plankton distributions in oceanic surface waters can be characterized by the spectra of concentrations obtained along ship transects. Such spectra are inevitably found to have a power-law form over horizontal scales ranging from 1 to 100 km (ref. 2). Phytoplankton have distributions similar to those of physical quantities such as sea surface temperature, with much less variability at shorter length scales. In contrast, zooplankton density may be almost as variable at short scales as long ones³. Distributions of this form are generated in a model of the turbulent stirring of coupled phytoplankton and zooplankton populations. The characteristic spatial patterns of the phytoplankton and zooplankton are a consequence of the timescales of their response to changes in their environment caused by turbulent advection.

The accuracy with which the ensemble-averaged dispersal of passive tracers in turbulent ocean flows may be described by eddy diffusion⁴ has encouraged the use of diffusion-based models to study the generation of patchiness in planktonic ecosystems^{5–8}. Diffusion causes variability to be transferred from small to large length scales, and so it is widely believed that “in order to be able to form a more finely grained pattern the zooplankton must have some

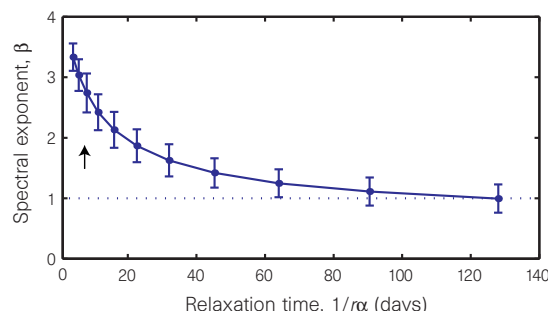


Figure 1 The phytoplankton carrying capacity relaxes towards a smoothly varying background value at a rate α , following equation 1. The power spectrum of a transect, $C(k)$, has the form $\tilde{C}(k) \propto k^{-\beta}$, where the spectral exponent, β , depends on the relaxation rate. Values are mean $\pm$ standard deviation of the exponents at the end of a high-resolution model run, calculated from 256 evenly spaced transects by linear regression of $\log(\tilde{C}(k))$ on $\log(k)$. With a relaxation time of $1/\alpha = 8$ d the spectrum has an exponent $\beta_0 = 2.74 \pm 0.32$ (arrow); this value of α is used in the text. The broken line marks the spectral exponent expected for a passive non-reacting tracer.

ability to resist turbulent diffusion"⁹. The long-standing history of this approach¹⁰, and the ease with which diffusion can be incorporated into differential equations, has led to the neglect of the fact that true diffusion within the oceanic surface layer is insignificant at larger than centimetre scales. At horizontal scales of between 1 and 100 km, only ensemble-averaged processes should be represented as diffusive. During any single realization, stirring by a turbulent flow causes variability to be transferred from large to small scales: under the action of turbulent advection a patch of tracer develops fine tendrils and filaments. The fine structure in the zooplankton distributions can then be generated by the transfer of variability from larger scales by stirring over the lifetime of the zooplankton. Models of the evolution of the mixed layer in a quasi-geostrophic flow suggest that at mid-latitudes approximately 10 days is required for the transfer of variance from the size of the largest turbulent features (~100 km) to the scale at which three-dimensional flow becomes important (~1 km)¹¹. This is less than the lifetime of larger zooplankton such as copepods, which typically require 25 days to reach adult stages¹². During their maturation time any large-scale

variation in the distribution of juvenile copepods will be stirred down to kilometre lengths.

The theory of turbulent transport has enabled phytoplankton patchiness to be understood in terms of both biological growth and physical advection^{13–17}, and dimensional analysis suggests that zooplankton–phytoplankton interactions may cause increased small-scale variability¹⁸. Although these theoretical approaches are remarkably powerful, they are difficult to apply to populations with anything beyond the simplest dynamics. In addition, the lack of a suitable representation of turbulent flow has restricted the direct incorporation of turbulent stirring into models of coupled plankton populations. Here the seeded-eddy model of a two-dimensional non-divergent flow¹⁹ is used to simulate turbulent advection (see Box 1). This model represents the turbulence as a sum of randomly distributed eddies of prescribed form. It is not a fluid-dynamic model, but with a judicious choice of parameters it provides a simple and easily implemented proxy for quasi-geostrophic turbulence. The domain of the model is a periodic square with sides of length L of 256 km. A number of independent, uniformly distributed water parcels are advected within this domain by the seeded-eddy flow. Each parcel carries a phytoplankton population density, P , and an adult zooplankton population density, Z , which represents a copepod species. While very large zooplankton, such as krill, may modify their distributions substantially by swimming²⁰. It is assumed here that the modelled zooplankton are truly planktonic, drifting with their respective water parcels.

In addition to these populations, each parcel has an associated carrying capacity, C , the maximum phytoplankton concentration attainable within that parcel in the absence of grazing. This carrying capacity may be assumed to represent the effect of a limiting nutrient, or to represent variations in mixed-layer depth. As a parcel moves through the domain, the carrying capacity continually relaxes towards a zonally varying background value, $C_0(x, y) = [1 - \cos(2\pi y/L)]/2$. This injects spatial variability into the model at the longest possible length scale. At the end of a model run the spatial distributions of the variables are obtained from a Delaunay triangulation of the parcel positions, by linear interpolation onto a regular grid²¹.

The dimensionless equations coupling the populations within a parcel are

$$dC/dt' = \alpha(C_0 - C) \quad (1)$$

$$dP/dt' = P(1 - P/C) - PZ \quad (2)$$

$$dZ/dt' = P(t' - \tau)Z(t' - \tau) - \delta Z^2, \quad (3)$$

where α is the relaxation rate of the carrying capacity; $t' = rt$ is the dimensionless time; r is the phytoplankton growth rate, which is chosen to be 0.5 d^{-1} ; τ/r represents the time taken for the zooplankton to mature; and δ is the zooplankton mortality. Phytoplankton growth is logistic and grazing is through a simple P – Z term: neither grazing saturation nor a minimum phytoplankton concentration at which grazing occurs are included. With $\tau = 0$ these equations reduce to a set previously used in the discussion of plankton patchiness⁵. In the absence of advection the model, at any position, has a single fixed point, $C = C_0$, $P = \delta C_0/(\delta + C_0)$, $Z = P/\delta$. This fixed point is stable for $\tau = 0$. If $C_0 \leq 1$, $\delta \geq 0.5$, as in these simulations, it remains stable when a non-zero maturation time is included.

The importance of response time for determining spatial structure is seen by comparing the distribution of the carrying capacity for different values of the rate parameter, α . When $1/\alpha$ becomes larger the carrying capacity adjusts more slowly to the turbulent advection. As a consequence the carrying capacity develops increased fine spatial structure and its spectra become flatter (Fig. 1). With $1/\alpha = 8 \text{ d}$, transects of the carrying capacity across

Box 1 The seeded-eddy model of a turbulent flow

In this implementation of the seeded-eddy model, the stream function is represented as a sum of N circular eddies

$$\psi(\mathbf{x}, t) = a \sum_{i=1}^N (\pm) R_i^2 e^{-\mathbf{x} \cdot \mathbf{x}_i(t)^2 / 2R_i^2}$$

where R_i and $\mathbf{x}_i = (x_i(t), y_i(t))$ specify the radius and centre position of each eddy, a is an overall calibration constant, and each eddy is included in the sum with a randomly assigned orientation. In these simulations the eddy centres are uniformly distributed over a 256-km square periodic domain, and move with the local velocity

$$\partial \mathbf{x}_i / \partial t = \mathbf{v}(\mathbf{x}_i)$$

where $\mathbf{v} = (-\partial \psi / \partial y, \partial \psi / \partial x)$. The turbulence is assumed to form a cascade between an upper radius, $R_{\max} = 64 \text{ km}$, appropriate for mid-latitude ocean turbulence and a lower radius, $R_{\min}$, the resolution limit of the model velocity field. Within this range the probability density of the radius is taken to be $p(R) \sim R^{-3}$, so the area covered by the eddies within each size interval, $\pi N \int_{R_{\min}}^{R_{\max}} p(r) dr$, is constant. The number of eddies, N , is chosen such that there are expected to be 16 eddies with a radius between 32 and 64 km. The spectra of the transverse and longitudinal velocity autocorrelation, taken from snapshots of the velocity field with $R_{\min} = 0.5 \text{ km}$ and $N = 87,536$, have an order-of-magnitude range of wavenumber where $E(k) \propto k^{-2.7 \pm 0.3}$, in close agreement with theoretical predictions²². Because the energy falls off rapidly at high wavenumber, a high-resolution model with $R_{\min} = 4 \text{ km}$ and $N = 1,360$ is sufficiently detailed. A low-resolution model, with $R_{\min} = R_{\max} = 64 \text{ km}$ and $N = 16$, is also used. The overall calibration is set to be $a = 0.14 \text{ d}^{-1}$ by requiring the r.m.s. velocity to be $|\mathbf{v}| \sim 10 \text{ cm s}^{-1}$. With these parameters the Lagrangian timescale of the seeded-eddy model is $T_L = 7 \pm 3 \text{ d}$, the Euclidean timescale is $T_E = 14 \pm 2 \text{ d}$ (both calculated from 16 300-day realizations), and the advective timescale $T_a = R_{\max}/|\mathbf{v}| = 7.4 \text{ d}$, in good comparison with experimentally determined values³⁰.

When the high-resolution model is used, 131,072 tracer parcels and their associated populations are followed, whereas the low-resolution model runs tracks only 8,192 parcels. Integration of both the biological and the physical components of the model is carried out using a simple finite difference scheme. The time step is restricted by the requirement that the parcels maintain a uniform distribution as they are advected. A time step of 0.1 d was found to be sufficiently small. The populations were initialized at their mean equilibrium values and then followed for 300 d. The model was programmed in FORTRAN and run on a Pentium 166-MHz computer, needing 4 s and 5 min for single low- and high-resolution time steps, respectively.

the model domain have spectra with similar exponents to the spectra of physical quantities such as sea surface temperature²², so the relaxation rate is set at $\alpha = 0.25$. At large relaxation times the spectral exponent becomes close to 1, the value expected for the spectrum of a passive non-reacting tracer in a two-dimensional turbulent flow¹⁸. This suggests that the turbulent transfer of variability to smaller scales is adequately represented by the seeded-eddy model.

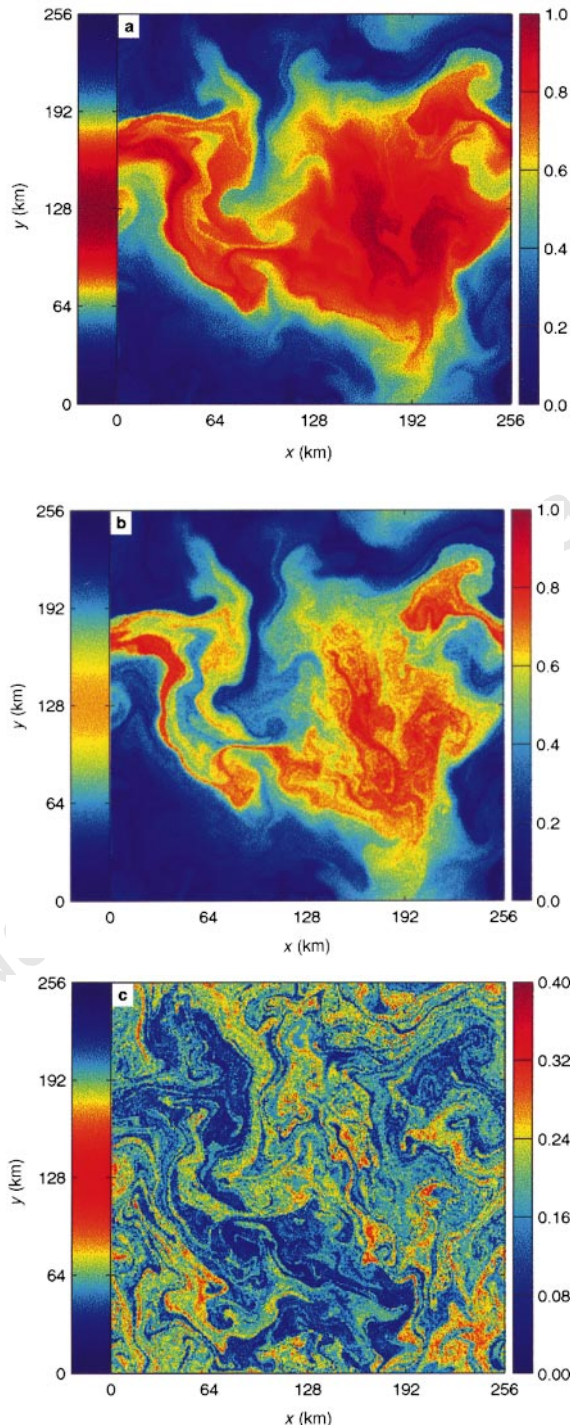


Figure 2 Snapshots at the end of a high-resolution model run. The model follows equations 1–3, with $\pi/r = 25$ d and $\delta = 2$, corresponding to a high P and low Z regime. **a**, Carrying capacity; **b**, phytoplankton; **c**, zooplankton. The strip at the left shows the zonally varying distributions the populations would have in the absence of advection while the bar on the right gives the values associated with the different colours. The distortion due to turbulent stirring is clearly visible.

With the inclusion of advection the continually changing carrying capacity prevents the populations within each parcel from reaching equilibrium. Despite the simplicity of the dynamics, a complex spatial pattern emerges by the end of the model run (Fig. 2). Because of the rapid phytoplankton growth rate, the phytoplankton distribution (Fig. 2b) is similar to the distribution of the carrying capacity (Fig. 2a). In contrast, the zooplankton population (Fig. 2c) has marked fine scale structure. This is clearly seen in a transect through the model domain (Fig. 3a), which simulates the data that would be collected from a ship. There is no coherence between the zooplankton and phytoplankton distributions at the larger length scales, and grazing causes the distributions to be negatively correlated at distances of less than ~ 10 km. As a consequence, the phytoplankton concentration has a spectrum that is slightly flatter than that of the carrying capacity, but steeper than the zooplankton spectrum (Fig. 3b). These results are in good agreement with observations, which find similar spectra, and which show a similar negative correlation between phytoplankton and zooplankton populations at shorter length scales^{2,3,23}.

An analysis of the populations obtained by integrating the model with a range of parameter values, and using a simplified representation of the flow, confirms that the relative slopes of these spectra are insensitive to both the precise values of the parameters and the details of the stirring process (Fig. 4). An increasing zooplankton maturation time leads to flatter zooplankton spectra, whereas a

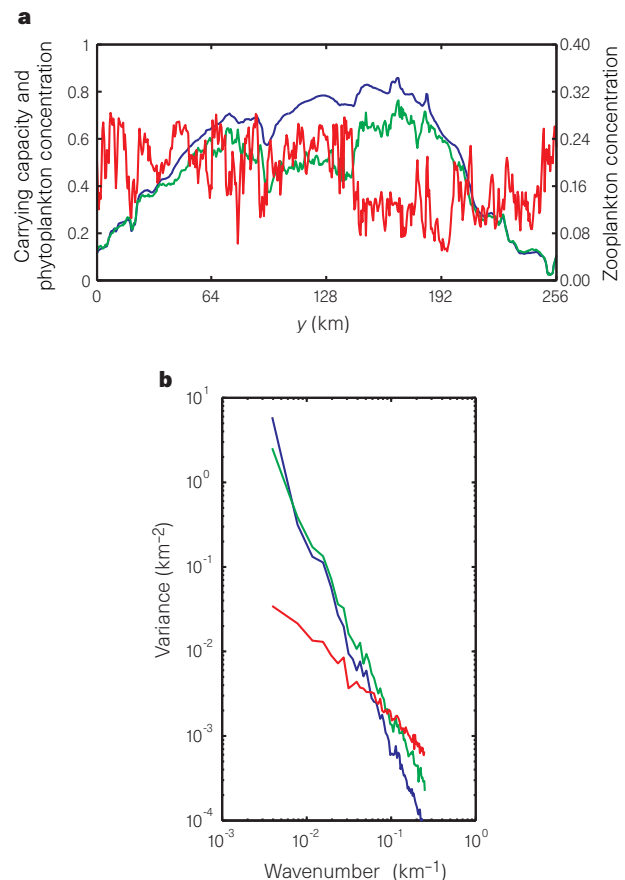


Figure 3 A representative transect and the corresponding spectra. Graphs show carrying capacity (blue), phytoplankton (green) and zooplankton (red). **a**, A transect through the snapshots in Fig. 2 (at $x = 128$ km). The simplicity of the underlying population dynamics is not apparent. **b**, The corresponding spectra have a power-law form over an order of magnitude range. The spectra from 256 evenly spaced transects are averaged to form those shown here. The spectral exponents (mean $\pm$ s.d.) of the populations at this time are $\beta_c = 2.5 \pm 0.20$, $\beta_p = 2.1 \pm 0.20$ and $\beta_z = 1.0 \pm 0.16$.

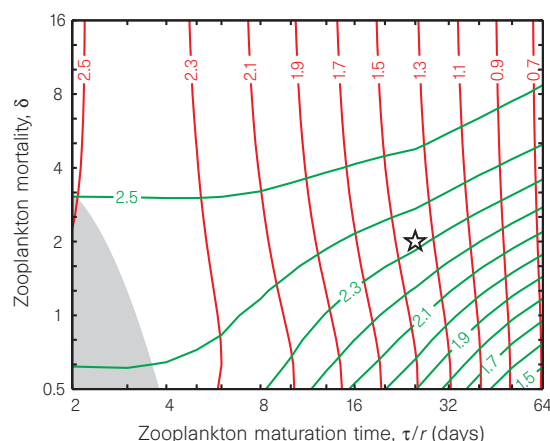


Figure 4 The dependence of the spectral exponents of the phytoplankton (green) and zooplankton (red) populations on the model parameters. The low-resolution model, including only the largest eddies, was run repeatedly with values of δ and τ indicated, but using the same velocity field. The contoured spectral exponents were obtained by averaging the spectra derived from snapshots of the populations at 10-day intervals between 160 and 300 model days. Within the shaded region, $\beta_z > \beta_p$; elsewhere, the zooplankton have a flatter spectra than the phytoplankton. The star marks the values of the parameters used in Figs 2 and 3.

larger zooplankton mortality leads to phytoplankton distributions with steeper spectra, because of the reduced grazing on the phytoplankton. Because the simplified physics, including only the largest eddies, is less efficient at transferring variability towards smaller length scales, the spectral slopes are steeper than the slopes generated by the model with a full turbulent cascade.

Detailed simulations, incorporating fluid-dynamic models of quasi-geostrophic turbulence, multi-compartment ecosystem dynamics, and seasonal forcing, have been attempted²⁴. These simulations find a similar result: that the zooplankton have a flatter spectra than the phytoplankton. Because of the complexity of the processes represented, the source of the fine-scale zooplankton structure in those models is unclear. The power of a simple model is to demonstrate how populations with these different distributions may occur without the need for any mechanism beyond their different response rates to changes in the environment caused by turbulent advection. Although predicting spectral form is a weak test of any theory²⁵, and the available evidence is perhaps ambiguous (not all data sets show flatly sloped zooplankton spectra²⁶, and similar results may arise as observational artefacts²⁷), this analysis suggests that zooplankton lifetime is an important determinant of their spatial pattern. A new observational sophistication, which enables zooplankton distributions to be tightly related to the physical properties of the surface waters²⁸, will allow a detailed testing of these relationships. □

Received 3 June; accepted 22 September 1997.

- Okubo, A. *Biomathematics* Vol. 10, *Diffusion and Ecological Problems: Mathematical Models* (Springer, Berlin, 1980).
- Mackas, D. L. & Boyd, C. M. Spectral analysis of zooplankton spatial heterogeneity. *Science* **204**, 62–64 (1979).
- Tsuda, A., Sugisaki, H., Ishimaru, T., Saino, T. & Sato, T. White-noise-like distribution of the oceanic copepod *Neocalanus cristatus* in the subarctic North Pacific. *Mar. Ecol. Prog. Ser.* **97**, 39–46 (1993).
- Okubo, A. Oceanic diffusion diagrams. *Deep Sea Res.* **18**, 789–802 (1971).
- Levin, S. A. & Segel, L. A. Hypothesis for origin of plankton patchiness. *Nature* **259**, 659 (1976).
- Wroblewski, J. S. & O'Brien, J. J. A spatial model of phytoplankton patchiness. *Mar. Biol.* **35**, 161–175 (1976).
- Steele, J. H. & Henderson, E. W. A simple model for plankton patchiness. *J. Plankton Res.* **14**, 1397–1403 (1992).
- Malchow, H. Nonequilibrium structures in plankton dynamics. *Ecol. Model.* **75/76**, 123–134 (1994).
- Parsons, T. R., Takahashi, M. & Hargrave, B. *Biological Oceanographic Processes*, 3rd edn (Pergamon, Oxford, 1984).
- Kierstead, H. & Slobodkin, L. B. The size of water masses containing plankton blooms. *J. Mar. Res.* **12**, 141–147 (1953).
- Klein, P. & Hua, B. L. The mesoscale variability of the sea surface temperature: An analytical and numerical model. *J. Mar. Res.* **48**, 729–763 (1990).

- Kiorboe, T. & Sabatini, M. 1995 Scaling of fecundity, growth and development in marine planktonic copepods. *Mar. Ecol. Prog. Ser.* **120**, 285–298 (1995).
- Denman, K. L. & Platt, T. The variance spectrum of phytoplankton in a turbulent ocean. *J. Mar. Res.* **34**, 593–601 (1976).
- Gower, J. F. R., Denman, K. L. & Holyer, R. J. Phytoplankton patchiness indicates the fluctuation spectrum of mesoscale oceanic structure. *Nature* **288**, 157–159 (1980).
- Lesieur, M. & Sadourny, R. Satellite-sensed turbulent ocean structure. *Nature* **294**, 673 (1981).
- Bennett, A. F. & Denman, K. L. Phytoplankton patchiness: inferences from particle statistics. *J. Mar. Res.* **43**, 307–335 (1985).
- Holloway, G. Eddies, waves, circulation and mixing: statistical geofluid mechanics. *Annu. Rev. Fluid Mech.* **18**, 91–147 (1986).
- Powell, T. M. & Okubo, A. Turbulence, diffusion and patchiness in the sea. *Phil. Trans. R. Soc. Lond. B* **343**, 11–18 (1994).
- Dyke, P. P. G. & Robertson, T. The simulation of offshore turbulent dispersion using seeded eddies. *Appl. Math. Model.* **9**, 429–433 (1985).
- Trathan, P. N., Priddle, J., Watkins, J. L., Miller, D. G. M. & Murray, A. W. A. Spatial variability of Antarctic krill in relation to mesoscale hydrography. *Mar. Ecol. Prog. Ser.* **98**, 61–71 (1993).
- Renka, R. J. Triangulation and interpolation at arbitrarily distributed points in the plane. *ACM Trans. Math. Software* **10**, 440–442 (1984).
- Deschamps, P. Y., Frouin, R. & Wald, L. Satellite determinations of the mesoscale variability of the sea surface temperature. *J. Phys. Oceanogr.* **11**, 864–870 (1981).
- Mackas, D. L., Denman, K. L. & Abbot, M. R. Plankton patchiness: biology in the physical vernacular. *Bull. Mar. Sci.* **37**, 652–674 (1985).
- Smith, C. L., Richards, K. J. & Fasham, M. J. R. The impact of mesoscale eddies on plankton dynamics in the upper ocean. *Deep Sea Res.* **43**, 1807–1832 (1996).
- Armi, L. & Flament, P. Cautionary remarks on the spectral interpretation of turbulent flows. *J. Geophys. Res.* **C90**, 11779–11782 (1985).
- Piontkovski, S. A. *et al.* Spatial heterogeneity of the planktonic fields in the upper mixed layer of the open ocean. *Mar. Ecol. Prog. Ser.* **148**, 145–154 (1997).
- Fasham, M. J. R. in *Spatial Pattern in Plankton Communities* (ed. Steele, J. H.) 131–156 (Plenum, New York, 1978).
- Gallagher, S. M., Davis, C. S., Epstein, A. W., Solow, A. & Beardsley, R. C. High-resolution observations of plankton spatial distributions correlated with hydrography in the Great South Channel, Georges Bank. *Deep Sea Res.* **43**, 1627–1663 (1996).
- Charney, J. G. Geostrophic turbulence. *J. Atmos. Sci.* **28**, 1087–1095 (1971).
- Middleton, J. F. Drifter spectra and diffusivities. *J. Mar. Res.* **43**, 37–55 (1985).

Acknowledgements. I thank P. Boyd, P. Sutton, C. Stevens and J. Sharples for critical reading of the draft manuscript, and R. Renka for software made available through the TOMS internet archive that was used to interpolate the Lagrangian variables onto a regular grid.

Correspondence and requests for materials should be addressed to E.A. (e-mail: e.abraham@niwa.cri.nz).

Collinear stimuli regulate visual responses depending on cell's contrast threshold

Uri Polat*, Keiko Mizobe*, Mark W. Pettet, Takuji Kasamatsu & Anthony M. Norcia

The Smith-Kettlewell Eye Research Institute, 2232 Webster Street, San Francisco, California 94115, USA

Neurons in the primary visual cortex are selective for the size, orientation and direction of motion of patterns falling within a restricted region of visual space known as the receptive field¹. The response to stimuli presented within the receptive field can be facilitated or suppressed by other stimuli falling outside the receptive field which, when presented in isolation, fail to activate the cell^{2–8}. Whether this interaction is facilitative^{3,4,7,9–12} or suppressive^{2,3,5,6,8–14} depends on the relative orientation of pattern elements inside and outside the receptive field. Here we show that neuronal facilitation preferentially occurs when a near-threshold stimulus inside the receptive field is flanked by higher-contrast, collinear elements located in surrounding regions of visual space. Collinear flanks and orthogonally oriented flanks, however, both act to reduce the response to high-contrast stimuli presented within the receptive field. The observed pattern of facilitation and suppression may be the cellular basis for the observation in humans that the detectability of an oriented pattern is enhanced by collinear flanking elements^{15–17}. Modulation of neuronal responses by stimuli falling outside their receptive fields may thus represent an early neural mechanism for encoding objects and enhancing their perceptual saliency.

* Present addresses: Institute for Vision Research, Ehad Ha'am 14, Rehovot 76105, Israel (U.P.); Department of Ophthalmology, Kyoto Prefectural University of Medicine, Hirokoji Kawara-machi, Kamigyo-ku, Kyoto 602, Japan (K.M.).

We recorded single-cell discharges extracellularly over a wide range of luminance contrasts to compare individual cells' contrast-response functions under four stimulus conditions: (1) a Gabor patch (discrete grating target) presented alone inside the classical receptive field (CRF) (Fig. 1a); (2) the CRF target plus two matching Gabor flanks that were well outside the CRF (Fig. 1b); and (3) the flanks alone (Fig. 1c). Orthogonally oriented flanks were also tested (Fig. 1d). The target-flank separation, centre-to-centre, was usually set at a 3–4 wavelength distance (the reciprocal of spatial frequency), because maximal perceptual facilitation occurs at this range^{15,16}. Typical cellular behaviour is depicted in Fig. 2a, in which response magnitude was plotted as a function of logarithmic target contrast. When the CRF was stimulated by the target alone, response magnitude increased as target contrast increased, with some saturation at the highest contrast tested (Fig. 2a, filled circles).

A major finding was revealed when the cell was stimulated with compound stimuli made of the target on the CRF and two high-contrast (80%) collinear flanks placed well outside the CRF. The

response was facilitated at low target contrasts and suppressed at high target contrasts (Fig. 2a, open circles). No modulatory effects were seen at intermediate target contrasts. When stimulated with the two flanks alone, the cell's activity remained at the background noise level (Fig. 2a, flanks alone), because stimulation of the CRF surround did not drive the cell. A similar pattern of modulation of the contrast-response function was obtained in another cell when the contrast-response function was measured using the swept-contrast method (see Methods). When tested with the target plus collinear flanks, we again found facilitation of the response at low target contrasts and suppression at relatively high contrasts (Fig. 2b). The flanks presented in isolation failed to elicit a response (Fig. 2b, flanks alone).

In a population of 96 cells (39 simple and 44 complex cells, and 13 undecided cells) recorded from five adult cats, 83 cells were tested with three contrast values or more. Most of the 83 cells showed modulation of either both facilitation and suppression (~34%), facilitation alone (~33%) or suppression alone (19%), when tested with various combinations of stimulus parameters. A fraction of

Table 1 Incidence of modulation types depending on stimulus types

		Orthogonal			Per cent
		Total	Facilitation	No effect	
Collinear	Total	150	8	104	38
	Facilitation	29	5	21	3
	No effect	99	3	78	18
	Suppression	22	0	5	17
	Per cent		5.3	69.3	25.3

Entries in each row and column indicate the number of responses ($n = 150$, from 34 cells) that showed the indicated modulation effect for the collinear and orthogonal flank stimuli, respectively. The asymptotic probability of the sample being drawn from a parent population that has identical response distributions for collinear and orthogonal flank stimuli was $P < 0.00003$ (Bowker's test for off-diagonal symmetry; $\chi^2 = 23.85$, d.f. = 3). This low P value indicates significant asymmetry in the distribution of facilitation and suppression between the collinear and orthogonal flanks.

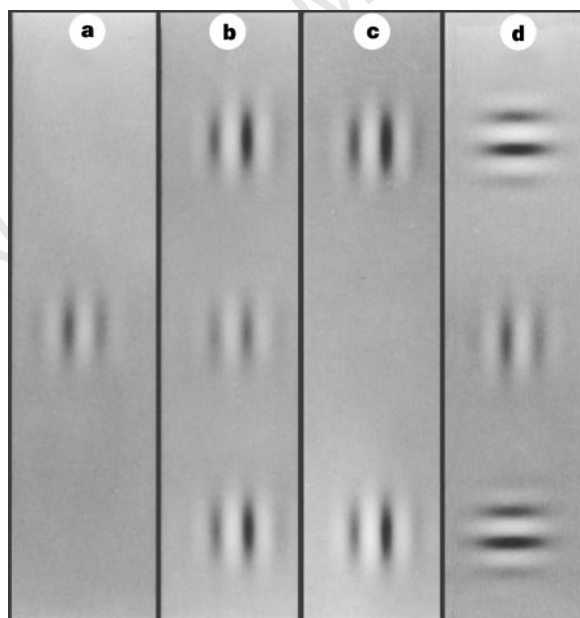


Figure 1 The stimuli were Gabor patches (Gaussian-weighted sinusoids) presented singly or in combination. **a**, Target patch optimally fitted to the classical receptive field (CRF) in its size, location, orientation and spatial frequency. **b**, Target patch presented concurrently with two flanking patches having the same properties as the target except for their contrast (collinearity test). The flanks were located well outside the CRF. **c**, Two flanking patches presented alone. **d**, Target and two flanks with orthogonal orientation. All other test parameters were the same as in **b**.

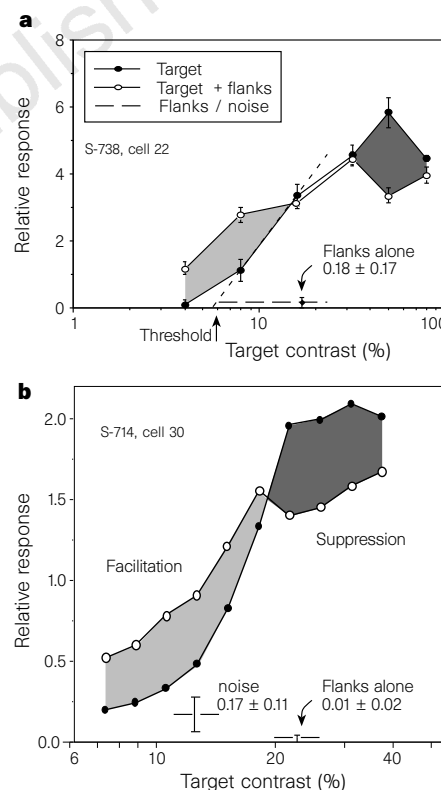


Figure 2 The comparison of contrast-response functions between target alone (filled circles) and target plus two collinear flanks (open circles) is exemplified in two single cells. The x-axis represents the target contrast as a percentage and the y-axis the cell's response magnitude in relative units of Fourier transforms, which varied depending on cells, and which were not normalized across cells. **a**, In the collinear configuration, the cell's response was facilitated at low target contrasts and suppressed at high target contrasts. The contrast threshold of this cell was ~6%, as estimated by linear extrapolation of the linear portion of the contrast-response curve to the noise level. The flanks alone elicited an insignificant response. Each vertical bar refers to one standard error of the respective mean (s.e.m.). This complex cell, recorded 986 μm from the cortical surface, had a moderate CRF size (3.2×2.4 deg). The average spike counts per run (9.7 s) at 8% contrast were as follows: 314.3 ± 59.2 (target alone), 488.5 ± 42.5 (target + flanks) and 269.4 ± 13.8 (flanks alone). **b**, Contrast response functions of another cell. The contrast-response function was measured by the swept-contrast method (see Methods). The cell response was facilitated at low contrasts and suppressed at intermediate and high contrasts. This simple cell, recorded at 1,126 μm from the cortical surface, had a relatively small CRF (1.0×0.8 deg).

cells showed no modulatory effects of the collinearly presented flanks (~14%).

In the 83 cells, we obtained 325 data points that satisfied a statistical test for significant modulation (target-flank additivity test, see Methods), covering three or more contrast values per cell. A population summary of the contrast-dependent modulation is shown for the collinear configuration in Fig. 3a. The distribution of the three modes of modulation (suppression, facilitation and no effect) was comparable between simple and complex cell populations (data not shown). Therefore, the two populations were combined using the first harmonic response for simple cells and the second for complex cells. For unclassified cells, the dominant harmonic was used. On average, two-fifths of the 325 data points accumulated from the 83 cells showed significant modulation, either suppression or facilitation, when tested with several target contrasts, the lowest being ~4% or lower, and the highest at 80%. The rest showed no modulation.

When the data were analysed separately within the modulated group, a suggestive trend emerged: facilitation was common at low contrasts and suppression was common at high contrasts (Fig. 3b). However, the trend was not significant (χ^2 test, $0.2 > P > 0.1$, d.f. = 5) owing to the presence of facilitation at relatively high target contrasts. For example, at 50% target contrast, facilitation was noted in over half of cases. Even at the highest target contrast (80%), over two-fifths of cases showed facilitation (Fig. 3b).

Psychophysical^{15,16}, visual-evoked-potential¹⁷ and modelling results^{11,12} have indicated that weak responses are facilitated by collinear surround stimuli. We therefore investigated whether facilitation at high target contrasts occurred in cells with moderate-to-high contrast thresholds, for example, in those cells that were relatively weakly activated despite the high target contrast. Estimating the cell's contrast threshold from the target contrast–response function (Fig. 2a; and see Methods), we re-evaluated the population behaviour of the above-mentioned modulatory effects in relation to the contrast threshold of individual cells.

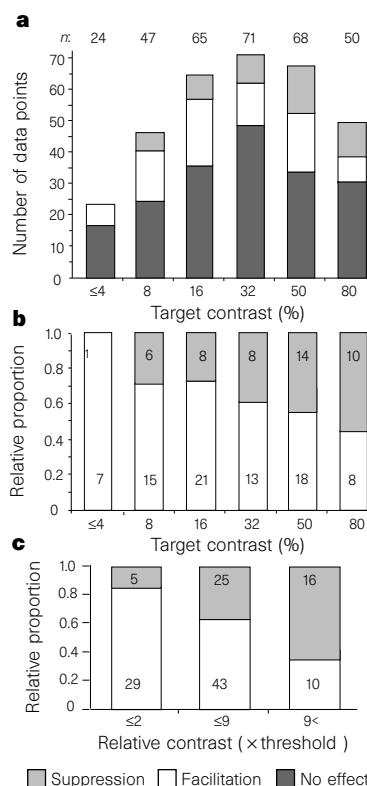
The frequency distribution of 'relative contrast' (multiples of threshold) had a large number of values around twice threshold, a

slowly declining number of measurements lying between 2 and 9 times threshold, followed by one-fifth of the sample scattered widely in a range larger than 9. Facilitative modulation was most common when the target contrast was less than or equal to twice the cell's contrast threshold. Suppression clearly dominated when the target contrast was more than nine times the threshold of individual cells (Fig. 3c). In an intermediate range of contrast (that is, more than twice, but equal to or less than nine times the threshold), the proportion of facilitative versus suppressive interactions fell in between. The effect of contrast—relative to the cell's own threshold—was highly significant (χ^2 test, $P < 0.001$, d.f. = 2).

Surround modulation can be orientation-related^{2–8,14}. We thus examined the effectiveness of orthogonally oriented flanks by accumulating 150 data points from 34 cells (including 24 cells from three additional cats) in which both collinear and orthogonal flanks were tested at the same contrast values. In this population, for the collinear configuration, significant response modulation was observed in 34% of the data points (29 and 22 responses for facilitation and suppression, respectively; Table 1). With the orthogonal configuration, response modulation occurred at about the same proportion (~31%), 38 responses showing suppression and 8 responses showing facilitation. Orthogonal flanks produced an increase in the relative proportion of suppressive interactions (from ~15 to ~25%) and a decrease in facilitation (from ~19 to ~5%). Both of these effects are consistent with an overall increase in suppression with the orthogonal configuration. The pattern of crossovers from one category to another, when collinear flanks were changed to orthogonal ones, significantly favoured suppression over facilitation (Bowker's test, $P < 0.00003$, $\chi^2 = 23.85$, d.f. = 3). Thus, suppression appears to be a substantially more general phenomenon. Facilitation, on the other hand, appears to be more specific to spatial configuration and contrast, consistent with recent psychophysical reports^{15,16}.

A cell's contrast–response function usually comprises three regions: an accelerating nonlinear region at low contrasts, a linearly increasing region at intermediate contrasts, and a saturation region at high contrasts¹⁸. With near-threshold collinear stimuli, facilitation

Figure 3 Population summary of contrast-dependent modulation for 325 data points, all of which satisfied the additivity test (see Methods), obtained from the 83 cells tested with three contrasts or more per cell. Data from simple and complex cells were combined using the first harmonic response for the former and the second for the latter. For unclassified cells, the dominant response was used. **a**, Frequency histogram of three types of effects; suppression (light grey), facilitation (white) and no effect (dark grey). The number of data points (each cell contributing more than once) in each of the three categories is shown for 6 values of target contrast, from the lowest, equal to or lower than 4%, to the highest, 80%. The total number of data points per contrast group is indicated at the top of each column. **b**, The proportion of facilitation and suppression is shown within the modulation group (128 data points). These data are based on 128 of 304 cases, in which contrast threshold estimates are available. The number of data points for either facilitation or suppression at each contrast range is indicated on the face of each column. The facilitative effect was common at lower target contrasts and the suppressive one at higher target contrasts. The trend was, however, not significant (χ^2 test; $0.2 > P > 0.1$, $\chi^2 = 9.21$, d.f. = 5). **c**, Population summary of contrast threshold-dependent modulation for the same 128 data points shown in **b**. The proportion of facilitation (white bars) compared with suppression (light grey) is shown for three groups of relative contrast, expressed as multiples of the contrast threshold of the individual cells. The facilitative modulation was dominant when the target contrast was less than or equal to twice threshold, whereas suppression predominated above 9 times the threshold value. Intermediate results were obtained for relative contrasts greater than twice, but less than 9 times the threshold. The effect of relative contrast is highly significant (χ^2 test; $P < 0.001$, $\chi^2 = 14.09$, d.f. = 2).



may prevail because of an accelerating nonlinear combination of subthreshold inputs from the flanks with weak inputs from the CRF. This integration may be based on network properties of cortical cells^{11,12}. Facilitation near threshold may occur as the result of additional noise generated by flank stimulation¹¹ or because of excitation acting unopposed by inhibition¹⁹.

As target contrast increases, the effect of the collinear flanks reverses, becoming predominantly suppressive. Neurophysiological^{20,21} and computational evidence^{11,12,22} has indicated that strong, optimal stimuli elicit excitatory and inhibitory postsynaptic potential sequences, whereas only excitatory ones are detectable with weak stimuli. This means that, at high contrasts, the relative contribution of inhibitory potentials to the cell's conductance increases. Inhibitory interneurons generate spikes at twice the rate of pyramidal cells and are also less adaptable²³, so the excitation–inhibition balance may be further shifted from saturation to suppression by additional subthreshold inputs from the flanks. Inputs from collinear flanks therefore appear to be combined with inputs from the CRF dynamically to regulate the cell's contrast–response function in the same way as has been proposed for thalamocortical afferents^{12,22}.

Surround facilitation and suppression in primary visuocortical cells were first observed to be contrast-dependent with the present stimulus conditions (discrete counterphasing Gabor patches)⁹, consistent with results from psychophysical experiments in which comparable stimuli were used^{15,16}. More recently, contrast-dependent surround effects were also observed when a central grating target was flanked by an annular surround^{10,14}. The pattern of facilitation and suppression obtained with our stimulus conditions is similar to that claimed by Toth *et al.*¹⁰, but differs from that reported by Levitt and Lund¹⁴. Our stimuli mainly yielded facilitation just above the cell's contrast threshold. In their experiments¹⁴, a high-contrast annulus consistently suppressed the response to a low-contrast target inside the CRF; however, it was not clear whether their low-contrast range was near each cell's contrast threshold. The physiological results of Levitt and Lund¹⁴ are similar to psychophysical results obtained by Cannon and Fullenkamp^{24,25}, who, using the centre-annulus configuration, found that the annulus usually suppressed the perceived contrast of the central target^{24,25}. Facilitation was noted only over a narrow range of relative contrasts between the centre and annulus²⁵. In short, neurophysiological results emphasizing facilitation⁷, such as we see here, and others emphasizing suppression¹⁴ complement each other, in that different configuration-specific interactions in the primary visual cortex may provide a cellular basis of psychophysical phenomena that are specific to their respective stimulus configurations.

Facilitative and suppressive centre–surround interactions may be organized differently in order to serve different functions. Facilitative interactions may be preferentially organized for collinear elements that comprise extended contours, both at moderate⁷ and low contrasts, as in our results. Suppression is a more general phenomenon, occurring across all orientations and over a broad region around the CRF^{5–9,13,14}. Nonspecific suppression may act to rescale the contrast–response function to maximize differential sensitivity in the face of high image contrast²⁶. In addition, suppression that attenuates the responses of highly sensitive cells may equalize the firing rates of all cells involved in coding the same contour. The neural basis for the contextual modulation of the contrast–response function may lie in the intrinsic network of cells connected horizontally across several hypercolumns^{27,28} or in feedback connections from higher visual areas beyond the primary visual cortex or both. Long-range lateral or feedback interactions may thus generate a second-order interaction field which constitutes an early neuronal basis of saliency, perceptual grouping and contrast scaling in vision.

Note added in proof: Sengpiel *et al.* (*Exp. Brain Res.* **116**, 216–228, 1997) briefly reported on collinear facilitation of a low-contrast

target by a high-contrast surround. □

Methods

Stimulus pattern. The contrast of Gabor patches (mean luminance of 40 cd m⁻²) was periodically reversed at a rate optimal to activate the cell (1.2–6.6 Hz). Target contrast was varied between 0.5 and 80% and that of two flanks was constant at either 50 or 80%, depending on the cell's contrast threshold for the target. The target size, spatial frequency, phase and orientation were optimally fitted to the classical receptive field (CRF). The centre-to-centre separation between the target and flanks was set at 3–4 wavelengths distance (that is, the reciprocal of spatial frequency). If any responses were suspected with flanks alone, the separation was increased until the flanks alone at the distance did not produce a measurable response. The spatial phase was always the same (in-phase) between the target and flanks.

Physiological recording. Five adult cats were prepared for terminal physiological recording of single-cell discharges from striate cortex, using our standard procedures^{9,13}. The cat was anaesthetized with a gas mixture of N₂O : O₂ : CO₂ = 75 : 22.5 : 2.5, supplemented by continuous i.v. infusion of ~2 mg kg⁻¹ h⁻¹ pentobarbital, and paralysed with continuous i.v. infusion of 10 mg kg⁻¹ h⁻¹ gallamine triethiodide. Recordings were made within 10 deg from the projection site of the area centralis in striate cortex throughout the cortical layers. The CRF of visually responsive cells was characterized first using conventional methods^{9,13}. Each cell was classified as simple or complex according to the original criteria of Hubel and Wiesel¹.

Data analysis and significance. Spike counts were first obtained from peristimulus time histograms in synchrony with contrast reversal. Next, spectrum analysis was carried out using Fourier transforms. Here, we focused our analysis mainly on the first (1F) and second harmonic (2F) components as they were dominant. Error estimates and significance tests ($P < 0.05$) of each harmonic component were performed by using the T_{circ}^2 statistic²⁹. For each measurement, a linear prediction of the sum of responses of target alone and the two flanks alone was compared with the measured response obtained with the three patches presented together (additivity test). When the measured response differed significantly, being either larger or smaller, from the prediction, the cell's response was considered to be modulated by the flanks. Otherwise, the cell's response was considered to be independent of the presentation of the flanks (no effect).

Contrast threshold estimation. The contrast threshold was extracted from the contrast–response function in two steps. An initial estimate was obtained by the sweep method³⁰, in which contrast was incremented in 10 equally spaced logarithmic values from low to high during a stimulation period of 10 s (Fig. 2b). The slope of the first 'linear' portion of the contrast–response curve was linearly extrapolated to the noise level. Then, stimulating at 4–6 fixed contrast values (Fig. 2a) around the initial estimate, we obtained another contrast–response curve that covered a narrower range of contrasts than before. Finally, threshold was determined by executing the same linear extrapolation on the second curve. The estimated threshold always falls to the left of the point at which the response rises above noise. Insignificant values were excluded from the threshold estimate and from all subsequent analyses.

Received 3 June; accepted 20 October 1997.

- Hubel, D. H., Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **160**, 106–154 (1962).
- Blakemore, C. & Tobin, E. A. Lateral inhibition between orientation detectors in the cat's visual cortex. *Exp. Brain Res.* **15**, 439–440 (1972).
- Maffei, L. & Fiorentini, A. The unresponsive regions of visual cortical receptive fields. *Vision Res.* **16**, 1131–1139 (1976).
- Nelson, J. I. & Frost, B. J. Intracortical facilitation among co-oriented, co-axially aligned simple cells in cat striate cortex. *Exp. Brain Res.* **61**, 54–61 (1985).
- Knierim, J. J. & Van Essen, D. C. Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. *J. Neurophysiol.* **67**, 961–980 (1992).
- Li, C.-Y. & Li, W. Extensive integration field beyond the classical receptive field of cat striate cortical neurons—classification and tuning properties. *Vision Res.* **34**, 2337–2355 (1994).
- Kapadia, M. K., Ito, M., Gilbert, C. D. & Westheimer, G. Improvement in visual sensitivity by changes in local context: parallel studies in human observers and in V1 of alert monkeys. *Neuron* **15**, 843–856 (1995).
- Sillito, A. M., Grieve, K. L., Jones, H. E., Cudeiro, J. & Davis, J. Visual cortical mechanisms detecting focal orientation discontinuities. *Nature* **378**, 492–496 (1995).
- Mizobe, K., Polat, U., Kasamatsu, T. & Norcia, A. M. Lateral masking reveals facilitation and suppression from the same single cells in cat area 17. *Assoc. Res. Vis. Ophthalm. Abstr.* **37**, S493 (1996).
- Toth, L. J., Rao, S. C., Kim, D.-S., Somers, D. & Sur, M. Subthreshold facilitation and suppression in primary visual cortex revealed by intrinsic signal imaging. *Proc. Natl Acad. Sci. USA* **93**, 9869–9874 (1996).

11. Stemmler, M., Usher, M. & Niebur, E. Lateral interactions in primary visual cortex: a model bridging physiology and psychophysics. *Science* **269**, 1877–1880 (1995).
12. Somers, D. C., Nelson, S. & Sur, M. Effects of long-range connections on gain control in an emergent model of visual cortical orientation selectivity. *Soc. Neurosci. Abstr.* **20**, 1577 (1994).
13. Kitano, M., Niiyama, K., Kasamatsu, T., Norcia, A. M. & Sutter, E. E. Retinotopic and nonretinotopic field potentials in cat visual cortex. *Vis. Neurosci.* **11**, 953–977 (1994).
14. Levitt, J. B. & Lund, J. S. Contrast dependence of contextual effects in primate visual cortex. *Nature* **387**, 73–76 (1997).
15. Polat, U. & Sagi, D. Lateral interactions between spatial channels: suppression and facilitation revealed by lateral masking experiments. *Vision Res.* **33**, 993–999 (1993).
16. Polat, U. & Sagi, D. The architecture of perceptual spatial interactions. *Vision Res.* **34**, 73–78 (1994).
17. Polat, U. & Norcia, A. M. Neurophysiological evidence for contrast dependent long-range facilitation and suppression in the human visual cortex. *Vision Res.* **36**, 2099–2109 (1996).
18. Albrecht, D. G. & Hamilton, D. B. Striate cortex of monkey and cat: contrast response function. *J. Neurophysiol.* **48**, 217–236 (1982).
19. Somers, D. C., Todorov, E. V., Siapas, A. G., Toth, L. J., Kim, D.-S. & Sur, M. A local circuit integration approach to understanding visual cortical receptive fields. *Cerebr. Cort.* (in the press).
20. Douglas, R. J., Martin, K. A. C. & Whitteridge, D. An intracellular analysis of the visual responses of neurones in cat visual cortex. *J. Physiol.* **440**, 659–696 (1991).
21. Hirsch, J. A. & Gilbert, C. D. Synaptic physiology of horizontal connections in the cat visual cortex. *J. Neurosci.* **11**, 1800–1809 (1991).
22. Douglas, R. J., Martin, K. A. C. & Whitteridge, D. A canonical microcircuit for neocortex. *J. Neur. Comp.* **1**, 480–488 (1989).
23. McCormick, D. A., Connors, B. W., Lighthall, J. M. & Prince, D. A. Comparative electrophysiology of pyramidal and sparsely spiny stellate neurons of neocortex. *J. Neurophysiol.* **54**, 782–806 (1985).
24. Cannon, M. W. & Fullenkamp, S. C. Spatial interactions in apparent contrast: inhibitory effects among grating patterns of different spatial frequencies, spatial patterns and orientations. *Vis. Res.* **31**, 1985–1998 (1991).
25. Cannon, M. W. & Fullenkamp, S. C. Spatial interactions in apparent contrast: individual differences in enhancement and suppression effects. *Vis. Res.* **33**, 1685–1695 (1993).
26. Heeger, D. J. Normalization of cell responses in cat striate cortex. *Vis. Neurosci.* **9**, 181–197 (1992).
27. Fitzpatrick, D. The functional organization of local circuits in visual cortex: insight from the study of tree shrew striate cortex. *Cerebr. Cort.* **6**, 329–341 (1996).
28. Gilbert, C. D. & Wiesel, T. N. Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. *J. Neurosci.* **9**, 2432–2442 (1989).
29. Victor, J. D. & Mast, J. A new statistic for steady-state evoked potentials. *Electroenceph. Clin. Neurophysiol.* **78**, 378–388 (1991).
30. Norcia, A. M., Clarke, M. & Tyler, C. W. Digital filtering and robust regression techniques for estimating sensory thresholds from the evoked potential. *IEEE Eng. Med. Biol.* **4**, 26–32 (1985).

Acknowledgements. We thank S. P. McKee, D. Sagi, C. W. Tyler and M. Usher for helpful comments, E. Schmidt for technical support, and K. Swenson and V. Vildavski for software development. This work was supported by the National Eye Institute, the SKERF and Kyoto Prefectural University of Medicine.

Correspondence and requests for materials should be addressed to T.K. (e-mail: takuji@skivs.ski.org).

Pathogenesis of two axonopathies does not require axonal neurofilaments

J. Eyer*, D. W. Cleveland†, P. C. Wong‡ & A. C. Peterson§

*INSERM CJF 97-08 and University of Angers, CHU, 49033 Angers, France

†Ludwig Institute for Cancer Research, and Departments of Medicine and Neuroscience, University of California, San Diego, La Jolla, California 92093, USA

‡Division of Neuropathology, Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

§Laboratory of Developmental Biology, Molecular Oncology Group, H-5, Royal Victoria Hospital, McGill University, 687 Pine Avenue West, Montreal, Quebec, Canada H3A 1A1

Neurofilaments are a major component of the axonal cytoskeleton and their abnormal accumulation is a prominent feature of the cytopathology encountered in several neurodegenerative diseases^{1–8}. Thus, an attractive and widely held model of pathogenesis involves the participation of disrupted neurofilaments as a common toxic intermediate^{9–13}. Here, in direct contrast to this hypothesis, we show that two neurodegenerative disease models in the mouse, dystonia musculorum (dt)^{14,15} and a superoxide dismutase 1 (SOD1)-mediated form of human motor neuron disease (amyotrophic lateral sclerosis, ALS)^{16,17}, progress with little or no abatement on a transgenic background in which neurofilaments are withheld from the axonal compartment¹⁸. By specifically excluding a necessary role for axonal neurofilaments, our observations redefine the components of the pathogenic pathway leading to axon disruption in these two degenerative diseases.

We previously derived transgenic mice that accumulate a neurofilament H(NFH)- β -galactosidase fusion protein within large projection neurons¹⁸. At low concentrations, this multivalent protein crosslinks neurofilaments in neuronal perikarya, limiting their export to axons. Despite the massive accumulation of neurofilaments in perikarya and the reduced radial growth of neurofilament-deficient axons, survival of many neuron populations is unaltered until an advanced age^{19,20}. This surprising observation indicated that a normally distributed neurofilament cytoskeleton is not required for neuron maturation, function or extended survival and brought into question the role played by the neurofilament aggregates encountered in diverse neurodegenerative diseases. Here, to determine if axonal neurofilaments play an essential role in the pathogenesis of two murine axonopathy models, we exploited the capacity of the NFHlacZ transgene to prevent axons from being invested with a neurofilament cytoskeleton.

In the autosomal recessive disease dystonia musculorum (dt), giant axonal swellings develop in sensory neurons. The dt gene encodes BPAG1^{15,21} and the neuronal isoform of this protein has both actin and intermediate filament-binding motifs suggesting that it may interact both with microfilaments and with neurofilaments¹⁴. Although cell bodies of sensory neurons are relatively preserved in affected mice, their axons develop swellings rich in neurofilaments (Fig. 1a, b) and degenerate, typically leading to death of affected animals by one month of age^{22–25}. To determine if the axonal neurofilament accumulations play a role in dt pathogenesis we derived homozygous dt/dt mice bearing the NFHlacZ transgene. Eleven litters were obtained and the expected ratio of dystonic and non-affected mice (26:76) was observed. Littermates of the four expected genotypes were analysed at 7, 14, 18 and 26 days of age. In all mice bearing the NFHlacZ transgene, regardless of their genotype at the dt locus, neurofilaments were found sequestered in neuronal cell bodies whereas axons lacked a neurofilament cytoskeleton (Fig. 1c, d). Axonal swellings were counted at three levels in cervical spinal cords from affected mice, with and without the NFHlacZ transgene, on day 14, 18 and 26. The average number

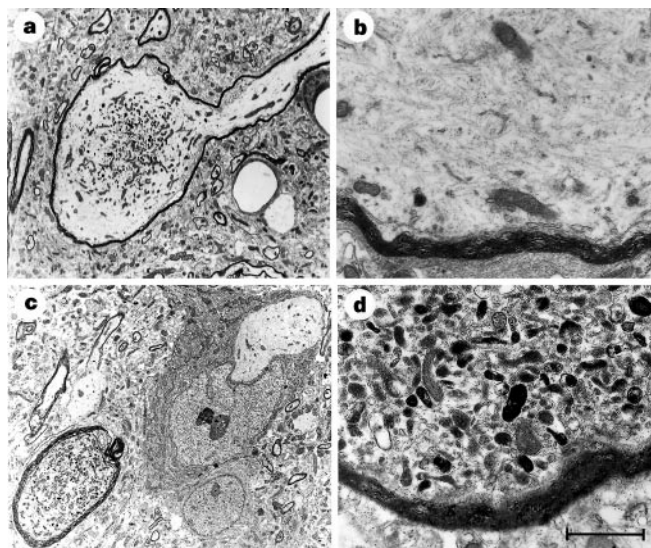


Figure 1 Electron micrographs of typical axonal swellings in myelinated fibres in the spinal cord grey matter of dt/dt mice. **a, b**, With the NFHlacZ transgene and **c, d**, without. In non-transgenic mice, neurofilament-filled axon swellings frequently reach more than 20 μ m in diameter. Within such swellings, organelles sequester in the middle whereas filament bundles course in random directions (**b**). In contrast, expression of the NFHlacZ transgene causes neurofilaments to aggregate and remain in the cell body as seen in the cell body profile in the upper right-hand corner of **c**. In dt/dt NFHlacZ transgenic mice, axonal swellings lack neurofilaments, are smaller and are filled with vesicles and mitochondria **d**. Scale bar, 7.6 μ m (**a, c**), 1.1 μ m (**b, d**).

of swellings per cross-section was 1.7 ± 0.7 and 18.0 ± 4.1 , respectively. When examined by electron microscopy, the rare swellings in *dt/dt* NFHlacZ mice were found to contain only vesicles and mitochondria with neither neurofilaments nor microtubules recognized at such sites (Fig. 1c, d).

Despite the low incidence of axonal swellings encountered in the transgenic samples, all affected (*dt/dt*) mice revealed a massive post-natal loss of dorsal root axons (Fig. 2). In affected non-transgenic mice, the population of myelinated fibres surviving in dorsal roots (lumbar levels 3, 4 and 5) at 14, 18 and 26 days was, respectively, 36%, 17% and 17% of the unaffected non-transgenic littermate values. Affected mice bearing the NFHlacZ transgene also demonstrated an early and massive loss of sensory axons but the proportion of longer-surviving axons was larger (45%, 40% and 30% of control values, respectively) (Figs 2d and 3). Thus, affected *dt/dt* mice without axonal neurofilaments continue to lose the majority of their sensory axons in a precipitous manner during the first month *ex utero*. Nonetheless, the modest attenuation in the rate of axon loss correlated positively with weight gain; at 26 days, unaffected non-transgenic (+/?, -/-) and transgenic (+/?, NFHlacZ/-) mice weighed $18.2 \text{ g} \pm 1.8$ and $17.2 \text{ g} \pm 1.5$, respectively, whereas affected non-transgenic mice (*dt/dt*, -/-) and affected transgenic mice (*dt/dt*, NFHlacZ/-) weighed $7.0 \text{ g} \pm 1.3$ and $14.4 \text{ g} \pm 3.0$, respectively.

In a few exceptional *dt/dt* NFHlacZ mice, lifespan extended beyond that typical of *dt/dt* mice. In one such 66-day-old mouse, the dorsal root axon population was 24% of the unaffected transgenic control values whereas in one 98-day-old mouse it had declined to 15%, similar to that observed in a 26-day-old affected non-transgenic mouse. Because the axon loss in these longer-surviving animals was only slowed and not prevented, we conclude that the neurofilament accumulation in axonal swellings is not an essential component of dystonia musculorum pathogenesis. Conversely, the transgene-mediated neurofilament maldistribution might stimulate neuronal stress responses or alter the transport capacity of axons, and as a result prolong the integrity of some

affected axons in *dt/dt* NFHlacZ mice. The obvious alteration in clinical course and the greater weight gain experienced in young *dt/dt* mice bearing the NFHlacZ transgene suggests that disproportionately large clinical benefits may be obtained by promoting relatively small changes in the size of surviving axonal populations.

Recently, several transgenic mouse models have been developed in which further neuronal populations are selectively prone to premature degeneration. Notably, *SOD1* transgenes containing mutations associated with inherited forms of ALS¹⁶ cause a dose- and age-dependent degeneration of motor neuron axons and cell bodies^{17,26}. One hypothesis linking *SOD1* mutations and motor axon degeneration is the enzyme-catalysed accumulation of nitrosylated tyrosine on prominent axon components such as neurofilaments²⁷. A damaged neurofilament cytoskeleton would then disrupt both fast and slow axonal transport leading to axonal degeneration²⁸. To investigate this hypothesis directly, we crossed transgenic mice bearing the ALS-linked *SOD1* mutant G37R to mice bearing the NFHlacZ transgene.

Typical of NFHlacZ mice, animals bearing both the NFHlacZ and G37R transgenes demonstrated perikaryal neurofilament aggregates and axons deficient in neurofilaments and rich in microtubules. G37R mutant *SOD1* mice from line 42 were analysed at 198 and 233 days of age. When fibres were examined at the level of the ventral root/spinal cord entry zone, most axon profiles were filled with multiple vesicles (Fig. 4a). Such vesicles, thought to arise from mitochondria¹⁷, were not observed at the mid-root level where

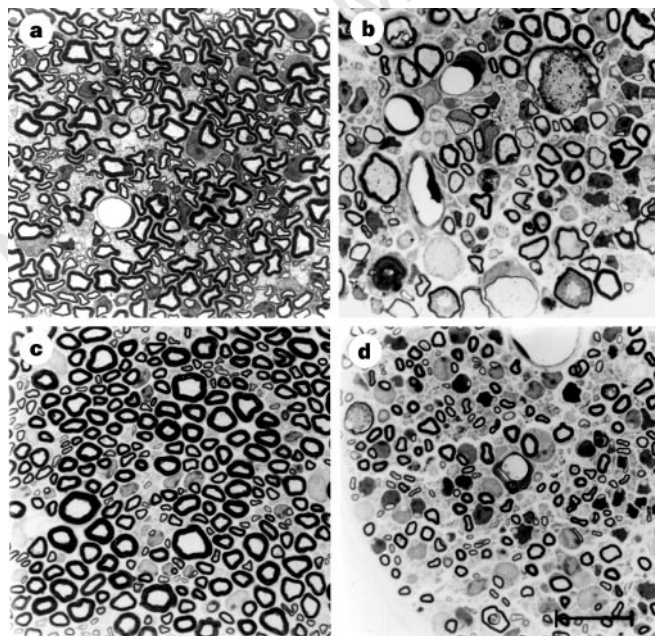


Figure 2 Cross-sections of dorsal roots from mice. **a, b**, Without the NFHlacZ transgene and **c, d**, with. L4 roots from 14-day-old normal control (**a**) and *dt/dt* littermate (**b**). Note the prominence of giant axons typical in *dt/dt* (**b**). L3 roots from 26-day-old NFHlacZ unaffected (**c**) and NFHlacZ affected (**d**) mice. In NFHlacZ transgenic mice, neurofilament-deficient axons achieve smaller than normal calibres. In **d**, although obvious axonal degeneration has occurred, note the continued sparing of many small- and medium-calibre axons and the paucity of large axonal swellings. Scale bar, $16.8 \mu\text{m}$.

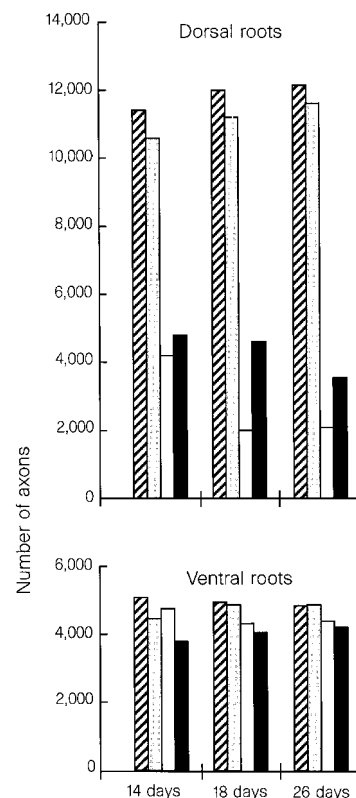


Figure 3 Axonal neurofilament deficiency does not prevent sensory axon degeneration in *dt/dt* mice. Littermates derived from crossing *dt/+* and *dt/+*, NFHlacZ/+ mice were killed at 14, 18 and 26 days of age and cross-sections of left and right dorsal and ventral roots (lumbar levels 3, 4 and 5) were prepared. Micrographs ($\times 787$) were assembled into montages containing profiles of complete root cross-sections and the number of myelinated fibres in each of the dorsal and ventral root populations counted. Unaffected non-transgenic (+/?, -/-), hatched bar. Unaffected transgenic (+/?, NFHlacZ/-), grey bar. Affected non-transgenic (*dt/dt*, -/-), white bar. Affected transgenic (*dt/dt*, NFHlacZ/+), black bar. Note the massive degeneration of dorsal root fibres in all *dt/dt* mice including those without axonal neurofilaments.

axons were frequently devoid of both mitochondria and vesicles. An indistinguishable combination of cytopathology (Fig. 4b) and axon degeneration was observed in mutant *SOD1* NFHlacZ animals that lack axonal neurofilaments.

To evaluate the extent of motor axon degeneration in *SOD1* mutant mice, with and without axonal neurofilaments, we evaluated both the morphology and the number of myelinated fibres in the middle of lumbar-level spinal roots. When *SOD1* mutant mice were near end-stage disease, the behavioural phenotype of their littermates bearing NFHlacZ alone was not distinguishable from wild-type animals. Surprisingly, little difference was observed between the total number of fibres present in NFHlacZ mice and *SOD1* mutant mice with or without the NFHlacZ transgene (Fig. 5a). Similarly, large-diameter profiles of fibres undergoing wallerian degeneration, recognized by the absence of both an axon and an intact compact myelin sheath, were encountered in mice of all three genotypes and the average number per root was not strikingly different, ranging between 10 and 20 and representing between 1.4% to 3.5% of the total myelinated fibre population.

The apparent paradox between the profound clinical deterioration of *SOD1* mutant mice and the unremarkable size of their ventral root axon population was solved on closer inspection of the fibres themselves. In all *SOD1* mutant mice, whether or not they carried the NFHlacZ transgene, a high density of small calibre and thinly myelinated fibres was observed (Fig. 5b, c). Frequently, such fibres were clustered and associated with Schwann cells demonstrating relatively large nuclear and cytoplasmic profiles; both features are consistent with axon sprouts acquiring myelin. As one damaged axon can elaborate multiple sprouts, the axon counts in both types of *SOD1* mutant mice overestimates the size of the axon population capable of maintaining muscle innervation. In contrast, as similar axon sprouts were not observed in mice bearing NFHlacZ alone (Fig. 5d), axon counts from these mice should accurately represent the intact axon population.

The similar quantitative and neuropathological features of ventral roots in *SOD1* mutant mice, with and without the NFHlacZ transgene, demonstrate that axonal neurofilaments play no essential role in the *SOD1*-mediated process leading to axonal degeneration. Further, the presence of axon sprouts in both types of *SOD1* mutant mice indicates that affected motor neurons support a regenerative response and that neither the perikaryal aggregates nor the deficiency of axonal neurofilaments has a marked effect on this capability.

The similar evolution of axonal cytopathology in *SOD1* mutant mice with and without the NFHlacZ transgene was also reflected in a similar disease course. The line 9 *SOD1* mice derived for this study

survived between 114 and 175 days ($N = 66$, $\bar{X} = 158 \pm 14.4$)¹⁷ and were indistinguishable from those also bearing the NFHlacZ transgene (range 151–173 days, $N = 11$, $\bar{X} = 166.6$, $P > 0.05$). In line 29 mice, which express low levels of the transgene-encoded *SOD1* protein, and in their doubly transgenic littermates, no clinical deficiencies were observed through one year of age. In line 42 *SOD1* mice, the majority survived less than 200 days ($N = 23$, $\bar{X} = 155.6 \pm 36.7$) and, as seen for line 9, the longevity of the

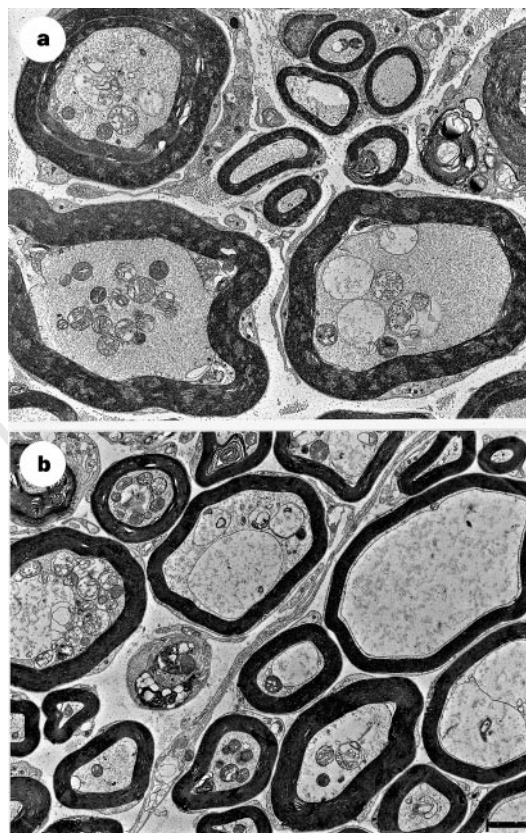


Figure 4 Mitochondrial degeneration occurs in axons of *SOD1* mutant mice, even in the absence of axonal neurofilaments. Electron micrographs of spinal root axons near the spinal cord/ventral root entry zone of line 42 *SOD1* transgenic mice at 198 days of age without (a) and with the NFHlacZ transgene (b). Most axons in both samples are filled with many membranous structures thought to arise from mitochondria. The cytopathology is similar in axons containing (a) or lacking (b) axonal neurofilaments. Scale bar, 1 μ m.

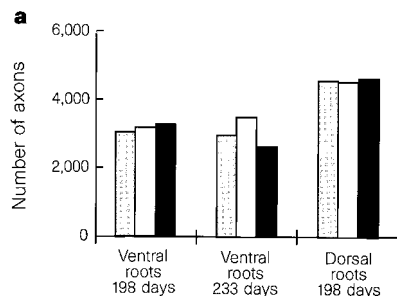
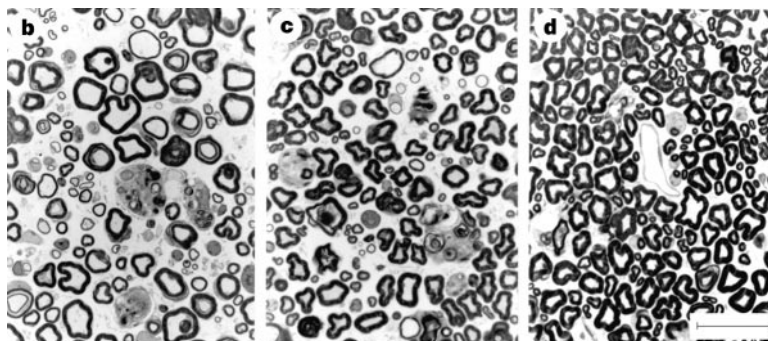


Figure 5 Axonal sprouts arise in ventral roots of *SOD1* mutant mice. a, Number of axon profiles present in cross-sections of ventral and dorsal roots of line 42 *SOD1* mice nearing end-stage disease and their littermates bearing only the NFHlacZ transgene. For ventral roots, the total number of myelinated axons present in left and right lumbar 3 and 4 levels is plotted (determined as described in Fig. 3). For dorsal roots, the left and right lumbar 4 level was counted. NFHlacZ transgenic mice without the *SOD1* transgene: grey bar, one animal at each age. *SOD1* mutant mice without the NFHlacZ transgene: white bar, one animal at each age.



Doubly *SOD1* and NFHlacZ transgenic mice: black bar, average from two animals at each age. The similar quantitative results from animals demonstrating such profound differences in clinical status prompted a qualitative assessment of the ventral root fibre populations. L4 ventral roots from *SOD1* mice without (b) and with (c) the NFHlacZ transgene contain multiple small-calibre and thinly myelinated fibres thought to be axonal sprouts whereas roots from a littermate bearing NFHlacZ alone (d) has only the small-calibre thickly myelinated fibres typical of mature neurofilament deficient axons. Scale bar, 23 μ m.

doubly transgenic population was not extended significantly ($N = 6$, $\bar{X} = 189.5 \pm 43.4$, $P > 0.05$). Thus, neither the proximal axonal cytopathology nor the clinical course of the SOD1 disease were modulated by either the NFHlacZ-mediated deficiency of axonal neurofilaments or the simultaneous accumulation of neurofilaments in motor neuron perikarya.

For the two mouse models of neuronal disease investigated here, neither initiation nor progression of pathology requires an axonal neurofilament cytoskeleton. This circumstance suggests that the neurofilament perturbations observed in these diseases arise as a passive response contributing little or nothing to the pathogenesis. It remains to be determined if and how far this conclusion can be generalized to human disease. The disparity between humans and rodents in both axon lengths and lifespan makes comparison of respective disease courses difficult. In addition, the possibility that interspecific amino-acid substitutions or additional genetic variation in neurofilament proteins may contribute to the pathogenesis of sporadic forms of ALS is the subject of debate^{29,30}. The resolution of these issues will be of particular significance in dissecting the pathogenic mechanisms underlying the many human late-life neurodegenerative conditions that exhibit abnormal neurofilament accumulations. □

Methods

Transgenic mice. NFHlacZ transgenic mice¹⁸ bear a fusion gene composed of mouse NFH, truncated in the fourth exon (EcoRV) and ligated in frame to bacterial LacZ. The *dt* allele used throughout this investigation resulted from insertion, within the *dt* gene, of a *hsplacZ* construct²³. Three lines of SOD1 mice (lines 9, 29 and 42), in which different levels of the human G37R mutant SOD1 protein accumulate¹⁷, were evaluated in this study. All animals were carefully monitored for onset and progression of disease. From the moment signs of weakness became apparent, a slow evolution to greater weakness occurred which was followed by a phase of rapid clinical deterioration. When the locomotion of affected animals became compromised they were considered to be in end-stage disease and were killed. Longevity results were compared by *T*-test.

Crosses were performed between (*dt*+/+, NFHlacZ/−) and (*dt*+/+, −/−) mice such that individual litters contained (*dt*/+, −/− and (*dt*/+, NFHlacZ/−) mice as well as unaffected transgenic and non-transgenic mice. For SOD1, littermates derived from crossing mutant SOD1/+ and NFHlacZ/+ mice were compared. The genotype of each animal was determined by polymerase chain reaction (PCR). For NFHlacZ primers (5′-AGGCTGCATCT CCA-GAAAAAGAAACC-3′) and (5′-TCATCATTAAGCGAGTGGCAACA-3′) were used. For *dt* we used: (5′-GATCCTGTGAGACGGGAGAAATGT-3′) and (5′-CACCTTGCTGAGGCAGGCTCTCC-3′), and for SOD1: (5′-CCAAGAT GCTTAACCTCTGTAATCAATGGC-3′) and (5′-CAGCAGTCTCATTGCCCA GGTCTCCAACATG-3′).

Electron microscopy and axon counts. Mice were anaesthetized with avertin (8 mg kg^{−1}) and perfused transcardially with 0.5% paraformaldehyde, 2.5% glutaraldehyde in 0.1 M phosphate buffer at room temperature (pH 7.4). Tissues were dissected and left in the same fixative overnight. After a brief rinse in 0.1 M phosphate buffer, the samples were post-fixed in 1% osmium tetroxide for 1 h, washed in 0.1 M phosphate buffer, and dehydrated in ethanol. The tissues were embedded in Epon, and sectioned for light or electron microscopy. Sections were stained with toluidine blue for light microscopy, and lead citrate and uranyl acetate for electron microscopy (JEOL 100 electron microscope).

To count the number of myelinated axons in each spinal root, a montage of the complete root was prepared using NIH Image software and an Olympus microscope (×100 objective) connected to a Macintosh PowerPC 7500 by a video camera. The number of myelinated axons was counted on each montage. To count the number of ventral root fibres undergoing wallerian degeneration, cross-sections were observed at ×630 with DIC optics using a Zeiss Axiophot microscope.

Received 18 June; accepted 4 November 1997.

1. Dickson, D. W. et al. Ballooned neurons in select neurodegenerative diseases contain phosphorylated neurofilament epitopes. *Acta Neuropathol.* **71**, 216–223 (1986).
2. Goldman, J. E. & Yen, S. H. Cytoskeletal protein abnormalities in neurodegenerative diseases. *Ann. Neurol.* **19**, 209–223 (1987).
3. Carden, M. J., Lee, V. M.-Y. & Schlaepfer, W. W. 2,5-Hexanedione neuropathy is associated with covalent crosslinking of neurofilament proteins. *Neurochem. Pathol.* **5**, 25–35 (1986).
4. Eyer, J., McLean, W. G. & Leterrier, J. F. Effect of a single dose of β,β′-Iminodipropionitrile in vivo on

the properties of neurofilaments in vitro: comparison with the effect of Iminodipropionitrile added directly to neurofilaments in vitro. *J. Neurochem.* **52**, 1759–1765 (1989).

5. Pollanen, M. S., Dickson, D. W. & Bergeron, C. Pathology and biology of the Lewy body. *J. Neuropathol. Exp. Neurol.* **52**, 183–191 (1993).
6. Hill, W. D., Lee, V. M.-Y., Hurtig, H., Murray, J. M. & Trojanowski, J. Q. Epitopes located in spatially separate domains of each neurofilament subunit are present in Parkinson's disease Lewy body. *J. Comp. Neurol.* **109**, 150–160 (1991).
7. Griffin, J. W. & Price, D. L. in *Experimental and Clinical Neurotoxicology* (eds Spencer, P. & Schaumburg, H. H.) 161–178 (Williams and Wilkins, Baltimore, 1980).
8. Carpenter, S. Proximal axonal enlargement in motor neuron disease. *Neurology* **18**, 841–851 (1968).
9. Lee, M. K. & Cleveland, D. W. Neurofilament function and dysfunction: involvement in axonal growth and neuronal disease. *Curr. Opin. Cell Biol.* **6**, 34–40 (1994).
10. Coté, F., Collard, J. F. & Julien, J. P. Progressive neuropathy in transgenic mice expressing the human neurofilament heavy gene: a mouse model of amyotrophic lateral sclerosis. *Cell* **73**, 35–46 (1993).
11. Lee, M. K., Marszalek, J. R. & Cleveland, D. W. A mutant neurofilament subunit causes massive, selective motor neuron death: implications for the pathogenesis of human motor neuron disease. *Neuron* **13**, 975–988 (1994).
12. Xu, Z. H., Cork, L. C., Griffin, J. W. & Cleveland, D. W. Increased expression of neurofilament subunit NF-L produces morphological alterations that resemble the pathology of human motor neuron disease. *Cell* **73**, 23–33 (1993).
13. Cleveland, D. W. et al. Mechanisms of selective motor neurons death in transgenic mouse models of motor neuron disease. *Neurology* **47**, 54–61 (1996).
14. Yang, Y. et al. An essential cytoskeletal linker protein connecting actin microfilaments of intermediate filaments. *Cell* **86**, 655–665 (1996).
15. Brown, A., Bernier, G., Mathieu, M., Rossant, J. & Kothary, R. The mouse dystonia musculorum gene is a neural isoform of bullous pemphigoid antigen 1. *Nature Genet.* **10**, 301–306 (1995).
16. Rosen, D. R. et al. Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* **362**, 59–62 (1993).
17. Wong, P. C. et al. An adverse property of a familial ALS-linked SOD1 mutation causes motor neuron disease characterized by vacuolar degeneration of mitochondria. *Neuron* **14**, 1105–1116 (1995).
18. Eyer, J. & Peterson, A. C. Neurofilament-deficient axons and perikaryal aggregates in viable transgenic mice expressing a neurofilament-β-galactosidase fusion protein. *Neuron* **12**, 389–405 (1994).
19. Plummer, J., Peterson, A. & Messer, A. Accelerated and widespread neuronal loss occurs in motor neuron degeneration (MND) mice expressing a neurofilament-disrupting transgene. *Mol. Cell. Neurosci.* **6**, 532–543 (1995).
20. Tu, P.-H. et al. Selective degeneration of Purkinje cells with Lewy body-like inclusions in aged NFHlacZ transgenic mice. *J. Neurosci.* **17**, 1064–1074 (1997).
21. Guo, L. et al. Gene targeting of BPAG1: abnormalities in mechanical strength and cell migration in stratified squamous epithelia and severe neurologic degeneration. *Cell* **81**, 233–243 (1995).
22. Duchon, L. W. Dystonia musculorum: an inherited disease of the nervous system in the mouse. *Adv. Neurol.* **14**, 353–365 (1976).
23. Kothary, R. et al. A transgene containing lacZ inserted into the dystonia locus is expressed in neural tube. *Nature* **335**, 435–437 (1988).
24. Campbell, R. M. & Peterson, A. C. An intrinsic neuronal defect operates in dystonia musculorum: a study of *dt*/+ ↔ +/+ chimeras. *Neuron* **9**, 693–703 (1992).
25. Sotelo, C. & Guenet, J. F. Pathologic changes in the CNS of dystonia musculorum mutant mouse: an animal model for human spinocerebellar ataxia. *Neuroscience* **27**, 403–424 (1988).
26. Gurney, M. E. et al. Motor neuron degeneration in mice that express a human Cu, Zn superoxide dismutase mutation. *Science* **264**, 1772–1775 (1994).
27. Beckman, J. S., Carson, M., Smith, C. D. & Koppenol, W. H. ALS, SOD and peroxynitrite. *Nature* **364**, 584 (1993).
28. Collard, J.-F., Cote, F. & Julien, J. P. Defective axonal transport in a transgenic mouse model of amyotrophic lateral sclerosis. *Nature* **375**, 61–64 (1995).
29. Figlewicz, D. et al. Variant alleles of the neurofilament heavy gene associated with amyotrophic lateral sclerosis. *Hum. Mol. Genet.* **3**, 1757–1761 (1994).
30. Vecchio, J., Bruijn, L., Brown, R. & Cleveland, D. Sequence variants in human neurofilament proteins: absence of linkage to familial amyotrophic lateral sclerosis. *Ann. Neurol.* **40**, 603–610 (1996).

Acknowledgements. We gratefully acknowledge the excellent technical assistance of T. Crossfield, K. Galloway-Kay, I. Tretjakoff and P. Valera. We also thank the 'Centre de Microscopie Electronique de l'Université d'Angers' and the Department of Neuropathology, MNI, for assistance in electron microscopy and image analysis. We are particularly grateful to J. Snipes, MNI, for helpful discussions. This investigation was supported through grants from the MRC and MDAC to A.C.P. and from the AFM to J.E. Travel support was provided by an INSERM/FRSQ award.

Correspondence and requests for materials should be addressed to A.C.P. (e-mail: Alan@devbiol2.molonc.mcgill.ca).

Blocking apoptosis prevents blindness in *Drosophila* retinal degeneration mutants

Florence F. Davidson* & Hermann Steller

Howard Hughes Medical Research Institute, Massachusetts Institute of Technology, 31 Ames Street, Cambridge, Massachusetts 02139, USA

Apoptosis is a gene-directed form of cell death that is essential for normal development and health. Yet abnormally high levels of apoptosis are linked to many degenerative diseases¹. Some important biochemical events in apoptosis have been identified², but the therapeutic utility of blocking cell death remains unclear. An important question in this regard is whether cells rescued from

* Present address: Department of Biochemistry, National Institutes of Health, National Cancer Institute, Bethesda, Maryland 20892, USA.

apoptosis can function. We have investigated the mechanism of cell death in two *Drosophila* mutant strains that exhibit age-related retinal degeneration. One of these mutations also occurs in humans, where it causes retinitis pigmentosa. We found that retinal cell death in *rdgC* and *ninaE^{RH27}/+* flies occurred by apoptosis and was blocked by eye-specific expression of the baculoviral cell survival protein p35. Most importantly, the mutant flies expressing p35 showed significant retention of visual function. The results demonstrate a therapeutic benefit of late-stage inhibition of apoptosis to flies, and suggest that similar results may be obtained in higher organisms.

Age related retinal degenerations, including retinitis pigmentosa (RP) are a leading cause of human blindness. In these diseases, initially functional photoreceptor cells of the eye are irreplaceably lost over time. There are two main mechanisms by which cells can die: necrosis, which is a lytic response to overwhelming stress, and apoptosis. Apoptosis is non-lytic, morphologically distinct, and its core elements are apparently conserved in all cells^{2–4,29}. Previous results indicate that cell death in rodent models of RP is apoptotic^{5–7}. About 30% of all autosomal dominant RP is caused by mutations in the photoreceptor protein rhodopsin⁸, and equivalent mutations in the rhodopsin gene of *Drosophila* photoreceptor R1–6 cells also cause dominant, age-related retinal degeneration^{9,10}. This suggests that the activation of cell death by such mutations occurs by a common mechanism, despite downstream differences in fly and human phototransduction pathways. The *Drosophila* strain *ninaE^{RH27}* carries a dominant rhodopsin 1 mutation, C200Y (ref. 10), whose equivalent in humans causes severe RP¹¹. Retinal degeneration in heterozygous *ninaE^{RH27}* flies begins at about 3

weeks of age¹⁰. The *rdgC³⁰⁶* strain carries a deletion of a serine/threonine protein phosphatase that appears to dephosphorylate light-activated rhodopsin^{12–14}. Retinal degeneration in homozygous *rdgC³⁰⁶* flies begins after 2 days in constant light. There is no known analogy to the *rdgC³⁰⁶* mutation in humans, but mutations near the phosphorylation sites of human rhodopsin are known to cause autosomal dominant RP¹⁵.

We examined the morphology of dead and dying cells in *ninaE^{RH27}/+* and *rdgC³⁰⁶* fly retinas to determine whether death was by apoptosis. Figure 1 shows representative electron micrographs of wild-type and *ninaE^{RH27}/+* fly retinas. Fly eyes are composed of symmetrically repeating cell clusters, or ommatidia. A typical cross-section of a wild-type ommatidium, in which the six peripheral photoreceptor cells (R1–6) can be seen enclosing a central photoreceptor cell (R7 here), is shown in Fig. 1a. Some of the pigment cells that separate each ommatidium are seen at the periphery. The morphology of a representative *ninaE^{RH27}/+* ommatidium after light-rearing for 8 weeks is shown in Fig. 1b. At this age, the cell bodies of all remaining R1–6 cells were small and the rhabdomere membranes reduced and disorganized. The cells showed the typical features of apoptosis: cytoplasmic condensation, nuclear chromatin condensation, and relatively normal looking mitochondria (Fig. 1d, e). Electron-microscopic examination of degenerating *rdgC³⁰⁶* retinas also revealed these features of apoptosis after 11 days in the light (not shown).

A common feature of apoptosis is the activation of cysteine proteases known as caspases¹⁶. The baculoviral cell survival factor p35 inhibits a broad spectrum of these^{17,18} and blocks apoptosis in insects^{19–22}, nematodes^{17,23} and mammalian cells^{24–26}. To test the

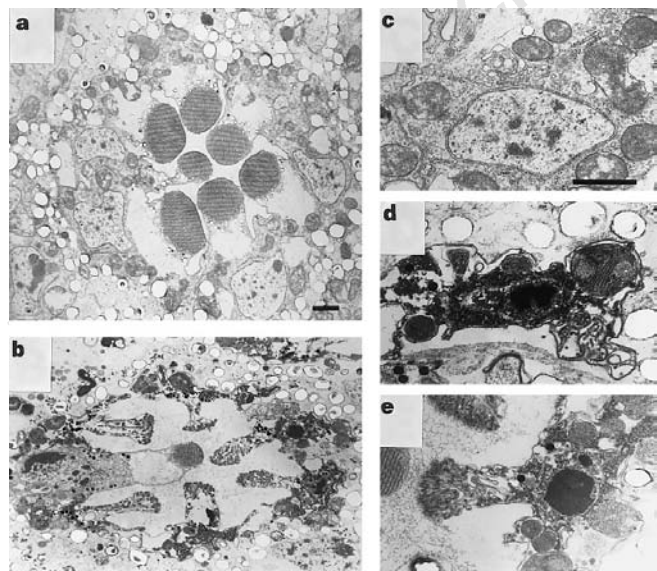


Figure 1 Photoreceptor cells in light-reared *ninaE^{RH27}/+* flies exhibit apoptotic morphology. **a**, Electron micrograph of a wild-type (Canton S) ommatidium. The membrane organelles at the centre of the ommatidium are the photoreceptor cell rhabdomeres, containing most of the cell's rhodopsin. **b**, An 8-week-old, light-reared *w/+; +; ninaE^{RH27}/+* fly ommatidium. Note the degeneration of the peripheral cells that express the mutant rhodopsin. The central photoreceptor cell, which expresses a different rhodopsin gene, looked normal here, but was degenerate in many ommatidia. **c**, A wild-type photoreceptor cell, with normal nuclear and mitochondrial morphology. **d**, **e**, High-magnification views of the cell bodies and subcellular organelles of the *w/+; +; ninaE^{RH27}/+* photoreceptor cells. Note the electron density of the cytoplasm and nuclei compared with the central photoreceptor cell, surrounding pigment cells, and wild-type cells (**a**). The normal spacing of the mitochondrial cristae indicates no lysis. Scale bars, 1 μ m.

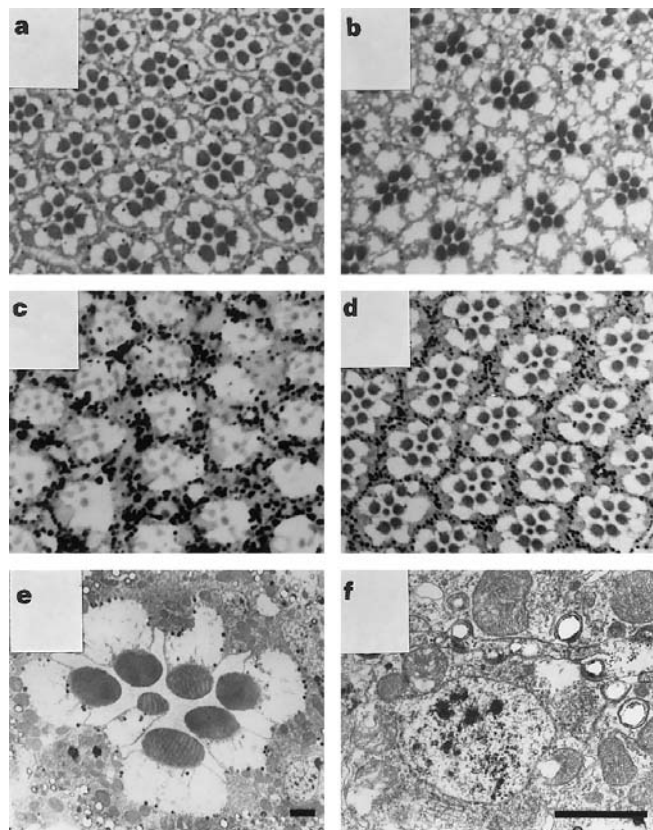


Figure 2 Retinal degeneration in *ninaE^{RH27}/+* flies is rescued by p35 expression. Light micrographs of retinal cross-sections of flies: **a**, wild type; **b**, dark-reared 4-week-old *w/+; +; ninaE^{RH27}/+*; **c**, light-reared 8-week-old *w/+; +; ninaE^{RH27}/+*; and **d**, light-reared 8-week-old *w/+; +; ninaE^{RH27}/IP[GMRp35 w^3-5]*. Scale bar, 15 μ m. **e**, **f**, Electron micrographs of an ommatidium (**e**) and cell bodies and organelles (**f**) show normal morphology in the light-reared 8-week-old *w/+; +; ninaE^{RH27}/IP[GMRp35 w^3-5]* fly retina. Scale bar, 1 μ m.

Table 1 Amplitudes of ERG response components in *rdgC* flies with and without p35.

Light exposure	Genotype	N	ERG component (mV)		
			On	Sustained	Off
+	wild type	7	0.8 ± 0.3	5.4 ± 1.2	1.4 ± 0.8
+	<i>rdgC</i>	3	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
+	<i>p35; rdgC</i>	6	0.6 ± 0.3	6.3 ± 1.7	0.8 ± 0.4
–	wild type	5	0.7 ± 0.4	7.0 ± 3.4	1.4 ± 0.4
–	<i>rdgC</i>	3	0.9 ± 0.4	8.2 ± 1.4	3.9 ± 1.2
–	<i>p35; rdgC</i>	3	1.6 ± 1.3	8.0 ± 2.5	2.4 ± 0.3

Amplitudes are ± s.e.m. Full genotypes are as indicated in Fig. 5.

effect of p35 on retinal degeneration, *ninaE^{RH27}* and *rdgC³⁰⁶* flies were crossed with flies carrying single inserts of the transgene P[GMRp35]²⁰, which confers eye-specific expression of p35 under the control of the GMR promoter²⁰. Figure 2 shows representative light micrographs demonstrating that retinal degeneration was blocked in *ninaE^{RH27}/+* flies by the presence of the GMRp35 transgene, supporting the conclusion that photoreceptor cell death in this mutant occurs by apoptosis. The retinæ of light-reared wild-type and young, dark-reared *ninaE^{RH27}/+* flies were indistinguishable (Fig. 2a, b), whereas those of light-reared *ninaE^{RH27}/+* flies 8 weeks old showed extensive loss of photoreceptor cell rhabdomeres and bodies (Fig. 2c). However, the light-reared *ninaE^{RH27}/P[GMRp35]* flies showed no degeneration at this age (Fig. 2d). Prevention of degeneration by p35 was also found when P[GMRp35] was inserted at a different site (not shown), indicating that the transgene's protective effect was not due to fortuitous disruption of a gene required for cell death. Electron micrographs of the *ninaE^{RH27}/P[GMRp35]* fly retina shown in Fig. 2d demonstrate that its photoreceptor cells and organelles were indistinguishable from those of wild-type flies even at this level of resolution (Figs 1a, c, 2e, f). We also found that eye-specific expression of p35 prevented degeneration in *rdgC* flies (Fig. 3). Representative light micrographs of dark- and light-reared *rdgC* fly retinæ show that degeneration followed the expected light-dependent pattern¹² (Fig. 3a, b). Yet light-reared P[GMRp35]; *rdgC* flies showed no degeneration, even at a later age (Fig. 3c). Antidromic illumination (not shown) also indicated no degeneration in P[GMRp35]; *rdgC* flies after 3 weeks in the light.

The ability of p35 to prevent retinal cell loss in these mutants allowed us to examine whether cells that would ordinarily commit to apoptosis can still function, despite the physiological stress that normally triggers death. To test for visual function in *ninaE^{RH27}/+* and *ninaE^{RH27}/P[GMRp35]* flies, assays of a behaviour called the 'walking optomotor response' were used²⁷. This assay measures the ability of a fly to discriminate and follow a rotating pattern of black and white stripes. Wild-type (Canton S) flies scored 91 ± 6 (Fig. 4), consistent with previous reports²⁷. Light-reared *ninaE^{RH27}/+* flies showed normal optomotor responses up to 3 weeks of age (not shown), but by 5 weeks had lost all visual responsivity (Fig. 4). The visual responses of 5-week-old dark-reared *ninaE^{RH27}/+* flies were not significantly different from those of the wild type, indicating that the defective copy of the rhodopsin gene did not by itself block visual function. Most importantly, 5-week-old *ninaE^{RH27}/P[GMRp35]* flies showed wild-type-like optomotor responses regardless of their light history.

Similar results were found when the effect of p35 expression on visual function was tested in *rdgC³⁰⁶* flies. In the optomotor response assay, *rdgC* flies gave highly variable responses even when reared in the dark, where no morphological degeneration occurs. This variation could be an effect of *rdgC* in the eye, or could be due to a function of the *rdgC* protein in the brain, where it is also expressed¹³. Electrophysiological assays were therefore used to test vision in *rdgC* mutant flies. Figure 5 and Table 1 show that, although dark-reared *rdgC* flies gave wild-type-like electroretinograms (ERGs), the light-reared *rdgC* flies showed no detectable ERG

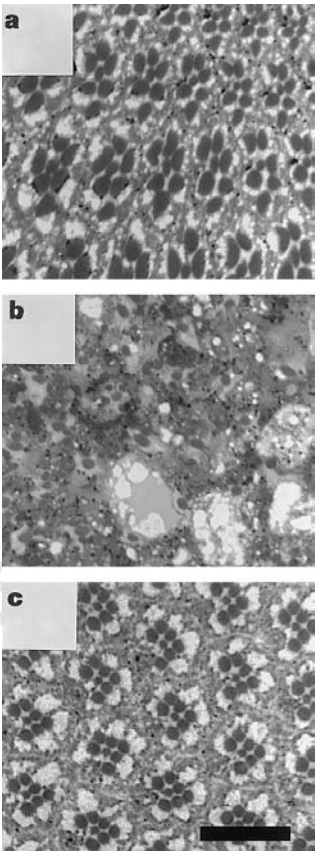


Figure 3 Retinal degeneration in *rdgC³⁰⁶* flies is rescued by p35 expression. Light micrographs of retinal cross-sections from flies: **a**, dark-reared 5-day-old +; +; *rdgC³⁰⁶, cu*; **b**, light-reared 5-day-old +; +; *rdgC³⁰⁶, cu*; and **c**, light-reared 11-day-old w; P[GMRp35 w]²⁻¹; *rdgC³⁰⁶, cu*. Scale bar, 15 μm.

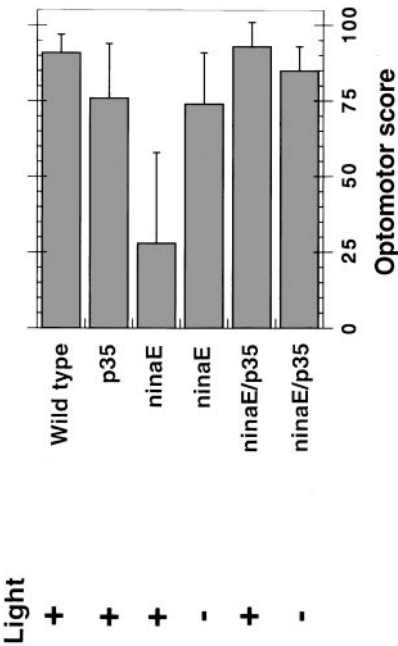


Figure 4 Optomotor responses in *ninaE^{RH27}/+* flies are rescued by p35 expression. The optomotor response scores (see text) are compared for flies raised for 5 weeks in either light (+) or dark (–). Wild type, Canton S flies; p35, transgenic flies expressing p35 in the retina (+; +; P[GMRp35 w]³⁻⁵); *ninaE*, *ninaE* mutants (w/+; +; *ninaE^{RH27}/+*); *ninaE*/p35, *ninaE* mutants expressing p35 in the retina (w; +; *ninaE^{RH27}/P[GMRp35 w]³⁻⁵*).

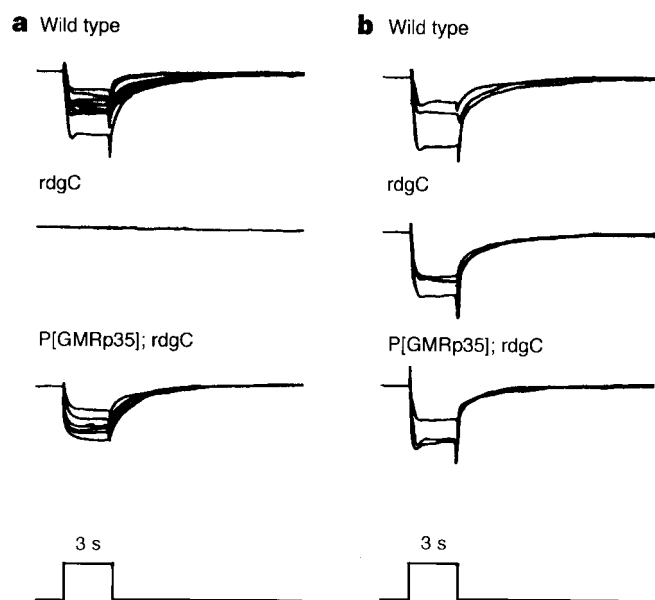


Figure 5 Electrophysiological responses in *rdgC*³⁰⁶ flies are rescued by p35 expression. All flies were aged to 7 days under the conditions indicated. **a**, Light-reared flies; **b**, dark-reared flies. The electroretinograms (ERGs) of wild type, *rdgC* and *P[GMRp35]; rdgC* flies record light-induced changes in potential that are a summed function of the photoreceptor cell depolarizations caused by activation of rhodopsin signalling. The mean amplitudes of these responses ($\pm$ s.e.m.) are shown in Table 1. Wild type indicates Canton S flies; *rdgC* indicates +; +; *rdgC*³⁰⁶, *cu* flies; *p35; rdgC* indicates *w*; *P[GMRp35 w⁺2-1; rdgC*³⁰⁶, *cu* flies.

responses by the same age. In contrast, light-reared *P[GMRp35]; rdgC* flies showed ERG amplitudes within the range of the wild type. The apparent 'on' rates of light-stimulated depolarization were slightly slowed in light-reared *P[GMRp35]; rdgC* flies than in their dark-reared siblings or other controls. The cause is unknown at present, but it may indicate some deceleration in downstream amplification²⁸ of the light signal. The visual signalling capacity of *P[GMRp35]; rdgC* flies eliminates the possibility that p35 protein prevents retinal degeneration by blocking rhodopsin function or expression. Moreover, this result is consistent with previous observations that the GMR promoter itself does not impede the expression of other *glass*-responsive genes.

In conclusion, we have determined that photoreceptor cell death in both *ninaE*^{RH27}/+ and *rdgC* light-dependent retinal degenerations proceeds by apoptosis and is inhibited by p35. This indicates that the activation of rhodopsin in these mutant backgrounds leads ultimately to caspase activation, and that caspase activation is required for degeneration. Our findings also made it possible to test whether the prevention of death would lead to restoration of function. It has been suggested that anti-apoptotic therapies may be useful for the treatment of various degenerative disorders¹. However, no information was previously available as to whether cells rescued from apoptosis under such conditions can serve a useful function. Measurable visual function was preserved in the structurally rescued *ninaE*^{RH27}/+ and *rdgC* mutant cells examined here, providing a strong rationale for further exploration of anti-apoptotic strategies in the treatment of degenerative disease. In particular, our results may provide a pointer for the effort to halt the progression of rhodopsin-mediated forms of RP. □

Methods

Fly strains and rearing conditions. Mutant *w*+/+; +; *ninaE*^{RH27}/+, *w*; +; *ninaE*^{RH27}/*P[GMRp35 w⁺3-5*], *w*; +/*P[GMRp35 w⁺2-1*]; *ninaE*^{RH27}/+ and +; *P[GMRp35 w⁺2-1*; *rdgC* flies were produced by standard genetic crosses from available fly stocks^{10,12,20}. Light-reared flies were kept at 25 °C in continuous

fluorescent light beginning 1–2 days after eclosion. Dark-reared flies were treated identically except for the exclusion of light from their environment. Canton S flies were used as wild-type controls.

Histology. Flies were decapitated under anaesthesia and the heads fixed in 2.5% glutaraldehyde in 0.1 M sodium cacodylate, pH 7.2, on ice for 30 min. OsO₄ was then added to a final concentration of 1% and fixation continued for 30 min. Glutaraldehyde and OsO₄ were removed and the heads fixed for 2 h in fresh 1% OsO₄ in the same buffer, rinsed with ice-cold water, stained in 1% aqueous uranyl acetate for 2 h, rinsed again with water, dehydrated through an ethanol series and propylene oxide, and embedded in Spurr's resin. Transverse semithin (0.5–1 μ m) sections were stained with methylene blue and toluidine blue for light microscopy. Transverse thin (50 nm) sections were post-stained with uranyl acetate and lead citrate for transmission electron microscopy at 80 kV.

Walking optomotor assays. The behavioural response of flies to a moving visual field was assayed as described²⁷. Normal flies attempted to stabilize their visual field in this assay by walking or turning in the direction of a rotating drum of black and white stripes. Behaviour was scored as the number of times each fly walked across a quadrant line in the same direction as the rotating stripes, divided by the number of times it walked in the opposite direction, multiplied by 100. Between 4 and 10 flies were tested per genotype, age and rearing condition. Every fly was tested in three successive trials, each consisting of a 1-min clockwise drum rotation followed by a 1-min anticlockwise rotation and a per-trial mean score calculated. Optomotor response scores were calculated as the average of the mean individual fly responses ($\pm$ s.e.m.) for that genotype, age and rearing condition. A score of 100 signifies that flies walk in the direction of visual field rotation 100% of the time. A score of 50% indicates that walking direction does not correlate with the direction of field rotation. Flies that lack functional photoreceptor cells and flies with impaired higher-order processing can both yield the latter score. Therefore, dark- and light-reared flies of the same age were tested to control for cognitive dysfunction unrelated to illumination history.

Electroretinograms. The electroretinograms (ERGs) of wild-type (Canton S), mutant +; +; *rdgC*³⁰⁶, *cu* and *w*; *P[GMRp35]2-1; rdgC*³⁰⁶, *cu* flies were obtained as described²⁷. A measuring electrode was inserted through the lens layer of one eye of each immobilized live fly, and a reference electrode was inserted through the cuticle at the back of the head. After dark adaptation for 3 min, the fly was illuminated with individual 3-s pulses of white light from a fibre-optic light source connected to a standard pulse generator. Electrode output was transmitted through a Grass P18 battery-operated preamplifier to a strip-chart recorder. All ERG measurements on mutant flies were bracketed by measurements on wild-type flies. Three recordings were taken per fly and 3–7 flies were tested per genotype, age and rearing condition. Each of the superimposed traces shows a sample ERG response obtained from one fly. Intra-fly means of the response component amplitudes were used to calculate the mean inter-fly response amplitudes ($\pm$ s.e.m.) in Table 1. Standard deviations of the mean of the intra-fly ERG response amplitudes were typically less than 10%.

Received 15 May; accepted 3 December 1997

- Thompson, C. B. Apoptosis in the pathogenesis and treatment of disease. *Science* **267**, 1456–1462 (1995).
- Steller, H. Mechanisms and genes of cellular suicide. *Science* **267**, 1445–1449 (1995).
- Hengartner, M. O. & Horvitz, H. R. Programmed cell death in *Caenorhabditis elegans*. *Curr. Opin. Genet. Dev.* **4**, 581–586 (1994).
- Jacobson, M. D., Weil, M. & Raff, M. C. Programmed cell death in animal development. *Cell* **88**, 347–354 (1997).
- Chang, G. Q., Hao, Y. & Wong, F. Apoptosis: final common pathway of photoreceptor death in rd, rds, and rhodopsin mutant mice. *Neuron* **11**, 595–605 (1993).
- Portera-Cailliau, C., Sung, C. H., Nathans, J. & Adler, R. Apoptotic photoreceptor cell death in mouse models of retinitis pigmentosa. *Proc. Natl Acad. Sci. USA* **91**, 974–978 (1994).
- Nickells, R. W. & Zack, D. J. Apoptosis in ocular disease: a molecular overview. *Optical. Genet.* **17**, 145–165 (1996).
- Humphries, P., Farrar, G. J., Kenna, P. & McWilliam, P. Retinitis pigmentosa: genetic mapping in X-linked and autosomal forms of the disease. *Clin. Genet.* **38**, 1–13 (1990).
- Colley, N. J., Cassill, J. A., Baker, E. K. & Zuker, C. S. Defective intracellular transport is the molecular basis of rhodopsin-dependent dominant retinal degeneration. *Proc. Natl Acad. Sci. USA* **92**, 3070–3074 (1995).
- Kurada, P. & O'Tousa, J. E. Retinal degeneration caused by dominant mutations in *Drosophila*. *Neuron* **14**, 571–579 (1995).
- Richards, J. E., Scott, K. M. & Sieving, P. A. Disruption of conserved rhodopsin disulfide bond by Cys187Tyr mutation causes early and severe autosomal dominant retinitis pigmentosa. *Ophthalmology* **102**, 669–677 (1995).
- Steele, F. & O'Tousa, J. E. Rhodopsin activation causes retinal degeneration in *Drosophila rdgC* mutant. *Neuron* **4**, 883–890 (1990).

13. Steele, F. R., Washburn, T., Rieger, R. & O'Tousa, J. E. *Drosophila* retinal degeneration C (rdgC) encodes a novel serine/threonine protein phosphatase. *Cell* **69**, 669–676 (1992).
14. Byk, T., Bar-Yaacov, M., Doza, Y. N., Minke, B. & Selinger, Z. Regulatory arrestin cycle secures the fidelity and maintenance of the fly photoreceptor cell. *Proc. Natl Acad. Sci. USA* **90**, 1907–1911 (1993).
15. Dryja, T. P. Rhodopsin and autosomal dominant retinitis pigmentosa. *Eye* **6**, 1–10 (1992).
16. Cohen, G. M. Caspases: the executioners of apoptosis. *Biochem. J.* **326**, 1–16 (1997).
17. Xue, D. & Horvitz, H. R. Inhibition of the *Caenorhabditis elegans* cell-death protease CED-3 by a CED-3 cleavage site in baculovirus p35 protein. *Nature* **377**, 248–251 (1995).
18. Bump, N. J. *et al.* Inhibition of ICE family proteases by baculovirus antiapoptotic protein p35. *Science* **269**, 1885–1888 (1995).
19. Clem, R. J. & Miller, L. K. Control of programmed cell death by the baculovirus genes p35 and iap. *Mol. Cell. Biol.* **14**, 5212–5222 (1994).
20. Hay, B. A., Wolff, T. & Rubin, G. M. Expression of baculovirus p35 prevents cell death in *Drosophila*. *Development* **120**, 2121–2129 (1994).
21. Grether, M. E., Abrams, J. M., Agapite, J., White, K. & Steller, H. The head involution defective gene of *Drosophila melanogaster* functions in programmed cell death. *Genes Dev.* **9**, 1694–1708 (1995).
22. White, K., Tahaoglu, E. & Steller, H. Cell killing by the *Drosophila* gene reaper. *Science* **271**, 805–807 (1996).
23. Sugimoto, A., Friesen, P. D. & Rothman, J. H. Baculovirus p35 prevents developmentally programmed cell death and rescues a ced-9 mutant in the nematode *Caenorhabditis elegans*. *EMBO J.* **13**, 2023–2028 (1994).
24. Rabizadeh, S., La Count, D. J., Friesen, P. D. & Bredeson, D. E. Expression of the baculovirus p35 gene inhibits mammalian neural cell death. *J. Neurochem.* **61**, 2318–2321 (1993).
25. Beidler, D. R., Tewari, M., Friesen, P. D., Poirier, G. & Dixit, V. M. The baculovirus p35 protein inhibits Fas- and tumor necrosis factor-induced apoptosis. *J. Biol. Chem.* **270**, 16526–16528 (1995).
26. Martinou, J. *et al.* Viral proteins E1B19K and p35 protect sympathetic neurons from cell death induced by NGF deprivation. *J. Cell Biol.* **128**, 201–208 (1995).
27. Rendahl, K. G., Jones, K. R., Kulkarni, S. J., Baguliy, S. H. & Hall, J. C. The dissonance mutation at the no-on-transient-A locus of *D. melanogaster*: genetic control of courtship song and visual behaviors by a protein with putative RNA-binding motifs. *J. Neurosci.* **12**, 390–407 (1992).
28. Ranganathan, R., Malicki, D. M. & Zuker, C. S. Signal transduction in *Drosophila* photoreceptors. *Annu. Rev. Neurosci.* **18**, 283–317 (1995).
29. Wyllie, A. H. Apoptosis and carcinogenesis. *Eur. J. Cell Biol.* **73**, 189–197 (1997).

Acknowledgements. We thank L. Smith and J. Hall for assistance and instruction in the measurement of optomotor responses and retinal ERGs; J. E. O'Tousa for *rdgC* and *ninaE^{RD17}* mutant strains; and J. Hall and H. R. Horvitz for critical reading of the manuscript. F.E.D. was a postdoctoral associate and H.S. is an investigator with the Howard Hughes Medical Institute.

Correspondence and requests for materials should be addressed to H.S. (e-mail: steller@wccf.mit.edu).

CD40 ligand on activated platelets triggers an inflammatory reaction of endothelial cells

Volker Henn^{*,†}, Joseph R. Slupsky^{†,‡}, Michael Gräfe[§], Ioannis Anagnostopoulos^{||}, Reinhold Förster[†], Gert Müller-Berghaus[‡] & Richard A. Krocze^{*}

^{*} Molecular Immunology, Robert Koch-Institute, 13353 Berlin, Germany

[‡] Haemostasis Research Unit, Max-Planck-Institut für Physiologische und Klinische Forschung, Kerckhoff-Klinik, 61231 Bad Nauheim, Germany

[§] Deutsches Herzzentrum, Augustenburger Platz 1, 13353 Berlin, Germany

^{||} Institute of Pathology, Klinikum Benjamin Franklin, Freie Universität Berlin, Germany

[†] Max-Delbrück-Center for Molecular Medicine, 13122 Berlin, Germany

[†] These authors contributed equally to this work.

CD40 ligand (CD40L, CD154), a transmembrane protein structurally related to the cytokine TNF- α , was originally identified on stimulated CD4⁺ T cells^{1–3}, and later on stimulated mast cells and basophils⁴. Interaction of CD40L on T cells with CD40 on B cells is of paramount importance for the development and function of the humoral immune system⁵. CD40 is not only constitutively present on B cells, but it is also found on monocytes, macrophages and endothelial cells, suggesting that CD40L has a broader function *in vivo*. We now report that platelets express CD40L within seconds of activation *in vitro* and in the process of thrombus formation *in vivo*. Like TNF- α and interleukin-1, CD40L on platelets induces endothelial cells to secrete chemokines and to express adhesion molecules, thereby generating signals for the recruitment and extravasation of leukocytes at the site of injury. Our results indicate that platelets are not only involved in haemostasis but that they also directly initiate an inflammatory response of the vessel wall.

Human platelets were analysed by flow cytometry for the expression of CD40L on the cell surface. CD40L was not detectable on unstimulated platelets, but activation of platelets by the agonist thrombin (monitored by staining for CD63) resulted in maximal expression of CD40L within one minute (Fig. 1); thereafter, CD40L levels gradually declined. CD40L was also released in the presence of collagen or ADP plus adrenaline as agonists (results not shown). These experiments demonstrate that preformed CD40L is stored in platelets and is translocated to the cell surface as part of the 'basic platelet reaction' in which rapid upregulation of CD63, P-selectin, and several other proteins on the cell surface is accompanied by the release of soluble mediators from intracellular granules^{6,7}.

Immunoprecipitation with the monoclonal antibody TRAP1 indicated that the 33K and 28K transmembrane forms of human CD40L⁸ are indistinguishable in platelets and activated CD4⁺ T cells (Fig. 2a). The ability of platelet CD40L to bind CD40 was verified by using a soluble CD40-immunoglobulin (CD40-Ig) chimaeric reagent for immunoprecipitation (Fig. 2b). Sequencing of the CD40L complementary DNA obtained from platelets and the megakaryoblastic lines MEG-01 and UT-7 by reverse transcription followed by polymerase chain reaction (RT-PCR) revealed identity to the CD40L sequence from T cells² (not shown). The CD40L molecule expressed on platelets is thus identical in structure and specificity to the CD40L expressed on activated CD4⁺ T cells.

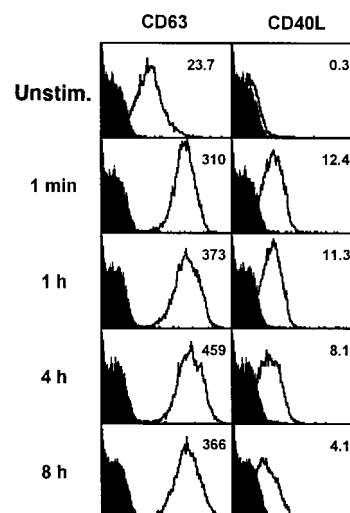


Figure 1 Expression of CD40L on activated platelets. Platelets were either left unstimulated or were stimulated with 0.2 U ml⁻¹ of human thrombin for the indicated times, fixed, and analysed by flow cytometry for expression of CD63 (or P-selectin; results not shown) and CD40L. The background signal is shown in black; inset numbers indicate the net geometrical mean of fluorescence of the cell population analysed. Unstim., unstimulated.

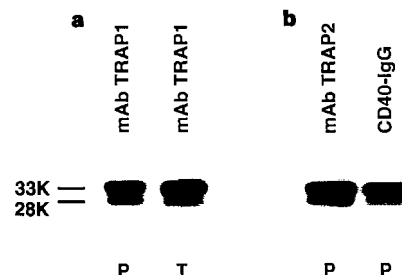


Figure 2 CD40L on platelets has the same structure and specificity as CD40L on activated CD4⁺ T cells. **a**, Immunoprecipitates of CD40L from platelets and obtained with TRAP1, or **b**, immunoprecipitates with a chimaeric CD40-Ig reagent (TRAP2 antibody was used for comparison) were analysed by western blotting with a CD40L-specific antiserum. P, platelets; T, T cells.

CD40 is constitutively expressed at low levels on the vascular endothelium of various organs, including spleen, lung and kidney¹⁰, suggesting that CD40 in the vascular system is a natural partner for CD40L on platelets. Ligand of CD40 on cultured human umbilical vein endothelial cells (HUVEC) was earlier shown to upregulate E-selectin (CD62E), VCAM-1 (CD106), and ICAM-1 (CD54)¹⁰⁻¹², all molecules that mediate the adhesion of neutrophils, monocytes and, later, lymphocytes to the inflamed vessel wall^{13,14}. When we co-incubated unstimulated platelets with HUVEC for 4 h, we observed a slight upregulation of all three adhesion molecules, probably reflecting a minimal preactivation of the platelets by the

experimental procedure (Fig. 3a). In contrast, platelets activated by physiological concentrations of human thrombin substantially upregulated E-selectin, VCAM-1 and ICAM-1 on HUVEC (Fig. 3a). This upregulation could be blocked to a large extent with the CD40L-specific antibody TRAP1 (E-selectin: $90 \pm 9\%$ s.e.m., $n = 5$; VCAM-1: $71 \pm 15\%$, $n = 5$; ICAM-1: $68 \pm 7\%$, $n = 5$), but not with the control antibody MOPC-21 (Fig. 3a). Thrombin alone had no significant effect (results not shown). Similar results were obtained with human arterial endothelial cells (not shown).

As our *in vitro* model probably only partially mimics the high density of CD40L available for interaction with CD40, when plate-

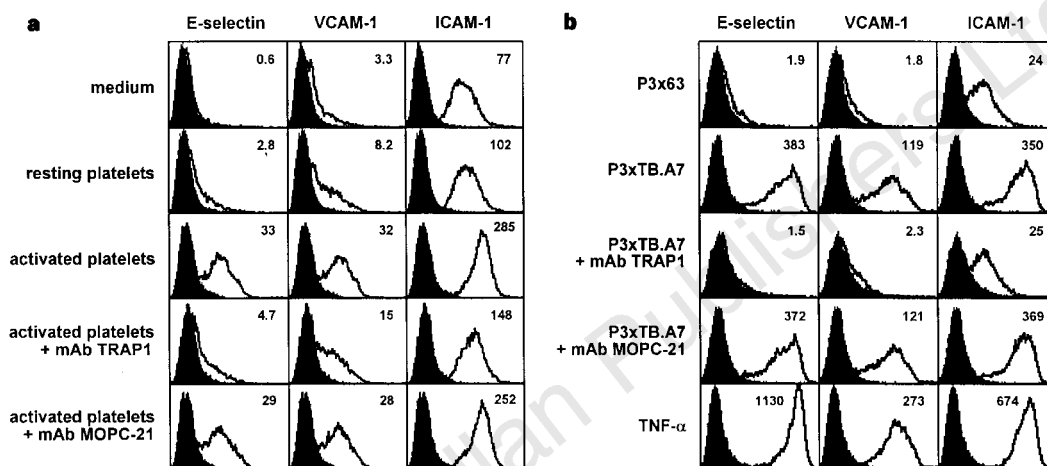


Figure 3 CD40L on activated platelets induces endothelial cells to express E-selectin, VCAM-1 and ICAM-1. HUVEC were co-incubated with **a**, medium alone, resting, or thrombin-activated platelets, or **b**, with P3x63 cells (wild-type control), or the CD40L-transfectant P3xTB.A7. After 4 h at 37 °C, in the presence or absence of mAb TRAP1 ($10 \mu\text{g ml}^{-1}$) or MOPC-21 ($10 \mu\text{g ml}^{-1}$, isotype control), the levels of E-selectin, VCAM-1 and ICAM-1 were determined by flow cytometry.

TNF-α was used at 100 U ml^{-1} . Paraformaldehyde fixation of the CD40L-transfectant before co-incubation gave similar results (not shown), thus excluding any influence of soluble factors. Addition of hirudin ($100 \mu\text{g ml}^{-1}$), a specific inhibitor of thrombin, to experiments with the CD40L transfectant excluded any effects of natural thrombin. Numbers indicate the net geometrical mean of fluorescence of the analysed cell populations.

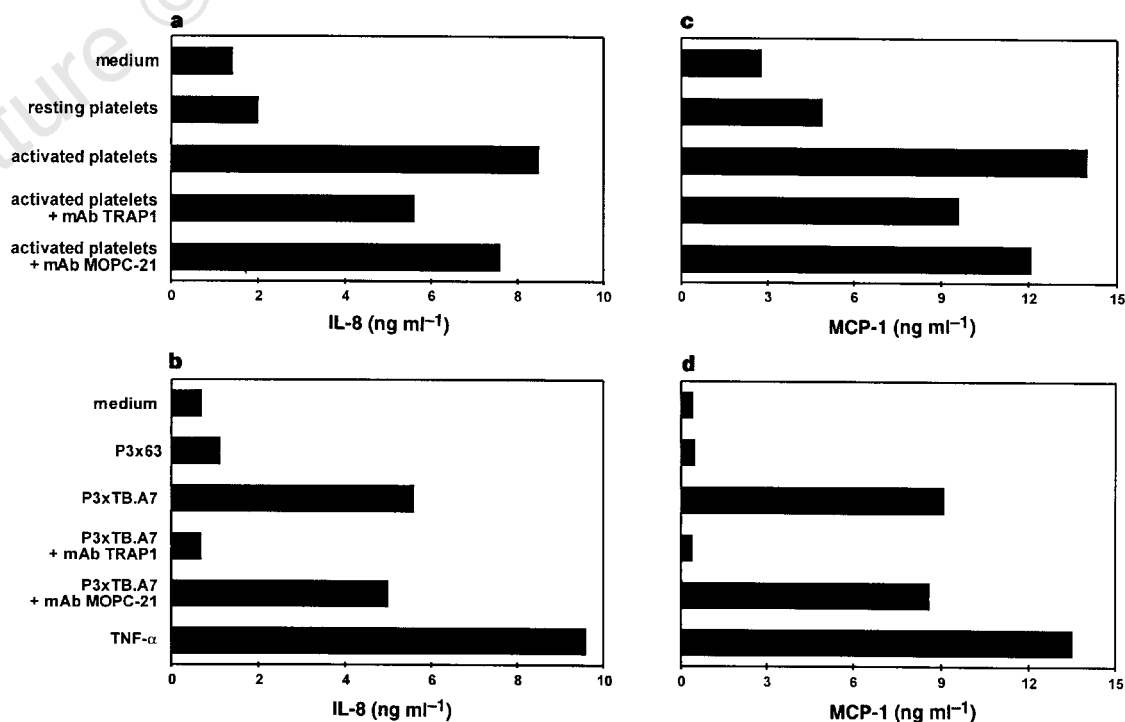


Figure 4 CD40L on activated platelets co-induces endothelial cells to secrete IL-8 and MCP-1. In the experiments described in Fig. 3, supernatants from co-incubations of HUVEC with platelets (**a, c**) or the CD40L-transfectant P3xTB.A7 (**b, d**) were collected after 4 h, and the levels of IL-8 (**a, b**) and MCP-1 (**c, d**) were

quantified using ELISA. Thrombin alone, at the concentration used, did not exert any significant effect.

lets adhere to and aggregate on the endothelium *in vivo*, we did additional experiments using the CD40L-transfected line P3xTB.A7⁸. This system enabled us to measure the isolated effect of membrane CD40L on the expression of the endothelial adhesion molecules. In contrast to earlier reports on HUVEC activation by CD40 (refs 10–12), we found that with the CD40L transfectant there was a dramatic upregulation of E-selectin, VCAM-1 and ICAM-1, which was fully blocked by TRAP1 (Fig. 3b). The magnitude of the effects achieved with membrane-bound CD40L was similar to those observed with TNF- α (and interleukin-1; not shown), which are two of the most potent mediators of inflammation (Fig. 3b).

Apart from upregulating adhesion molecules, endothelial cells react to inflammatory stimuli by secreting various chemokines, including interleukin (IL)-8, the principal chemoattractor for neutrophils, and MCP-1, which recruits and activates monocytes^{15,16}. Both chemokines can be used as indicators of the endothelial inflammatory reaction, because they are not stored in

platelets. After co-incubating activated platelets with HUVEC for 4 h, there was a substantial secretion of IL-8 and MCP-1 (Fig. 4a, c). This secretion was partially blocked by TRAP1 (IL-8: $39 \pm 19\%$ s.e.m., $n = 5$; MCP-1: $37 \pm 15\%$ s.e.m., $n = 3$), indicating that interaction of platelet CD40L with CD40 on endothelial cells contributes to the secretion of both chemokines by endothelial cells. Using the P3xTB.A7 transfectant once more, secretion of IL-8 and MCP-1 reached 50–75% of the response observed with TNF- α (Fig. 4b, d) and IL-1 (not shown).

To determine CD40L expression on activated platelets *in vivo*, we examined frozen tissue sections of fresh thrombi formed during the course of a vessel injury or on atherosclerotic plaques. The method used specifically allowed immunohistological detection of CD40L on the platelet surface (Fig. 5a, b). In all fresh thrombi analysed ($n = 5$), CD40L was expressed on a large proportion of platelets (40–80%) in areas in which densely packed platelets had not yet formed an amorphous mass (Fig. 5c–e). CD40L was also expressed on platelets directly adhering to the vessel endothelium (Fig. 5f). These results indicate that CD40L is expressed on platelets under conditions of intravascular thrombus formation *in vivo*.

Our findings alter the current T-cell-centric paradigm of major histocompatibility complex-restricted CD40L function *in vivo*. As $\sim 1 \times 10^{12}$ platelets in the circulation bear preformed CD40L molecules, the functional data obtained to date on the activation of endothelial cells, monocytes, inflamed fibroblasts and smooth-muscle cells by CD40 need to be re-evaluated. When platelets adhere to and become activated on dysfunctional endothelium¹⁷, or are triggered by contact with collagen upon extravasation, platelet CD40L immediately links haemostasis to the vascular inflammatory system¹⁸. Likewise, platelet CD40L is likely to trigger cells trapped in the thrombus, because monocytes secrete IL-1, TNF- α , IL-6, IL-8 and MIP-1 α upon engagement of CD40 (refs 19–21). Observations of IL-1 activity on stimulated platelets^{22,23} support this concept. The generation of inflammatory signals by platelets may thus occur following acute mechanical damage of the endothelium, in an infection of the vascular system, and also in the pathogenesis of atherosclerosis and vascular infarction, in which monocytes and platelets have pre-eminent roles²⁴. □

Methods

Platelet preparation and activation. Citrated platelet-rich plasma, to which EDTA (2 mM) and prostaglandin E1 ($1 \mu\text{g ml}^{-1}$; Sigma) were added, was centrifuged over a discontinuous albumin density gradient (25% over 34% endotoxin-free BSA (Sigma) in Tyrodes/HEPES buffer (138 mM NaCl, 2.9 mM KCl, 1 mM MgCl_2 , 1 mM glucose, 0.5 mM NaH_2PO_4 , 20 mM HEPES, pH 7.4)) at 750 g for 30 min. Platelets at the interphase were washed in Tyrodes/HEPES buffer, 0.3% BSA, by gel filtration on Sepharose-2B (Pharmacia). For time-course experiments, 2×10^8 platelets were incubated with 0.2 U ml^{-1} human thrombin (Boehringer Mannheim) for various times, fixed with 1% paraformaldehyde, washed, and analysed by flow cytometry.

Culture and activation of HUVEC. HUVEC were isolated from umbilical cords using collagenase and cultured in endothelial cell growth medium (PromoCell, Heidelberg), with endothelial cell growth supplement (PromoCell); the second passage was only in growth medium. For co-incubation experiments (4 h at 37°C), 3×10^8 platelets were activated by human thrombin (0.2 U ml^{-1}) and brought into close apposition with confluent HUVEC ($\sim 1 \times 10^6$ per plate) by centrifuging the plates at 750g for 2 min. P3xTB.A7-transfectant⁸ or P3x63 cells (both at 5×10^6), or TNF- α (100 U ml^{-1} ; Biomol, Hamburg) were added without centrifugation. TRAP1 (ref. 8) or MOPC-21 (isotype control) were used at $10 \mu\text{g ml}^{-1}$. After co-incubation, HUVEC were treated with 20 mM HEPES, pH 7.4, 100 mM NaCl, 10 mM EDTA, 0.5% BSA, for 20 min at 4°C and 10 min at 37°C , dislodged by vigorous pipetting, and analysed by flow cytometry. IL-8 or MCP-1 concentrations were measured in the supernatants from co-incubation experiments using specific ELISA reagents (R&D Systems).

Flow cytometry. Platelets were stained with biotinylated TRAP1 (ref. 8; IgG1, CD154), 1787 (CD63; Chemicon), or MOPC-21 (IgG1 isotype; Sigma) and

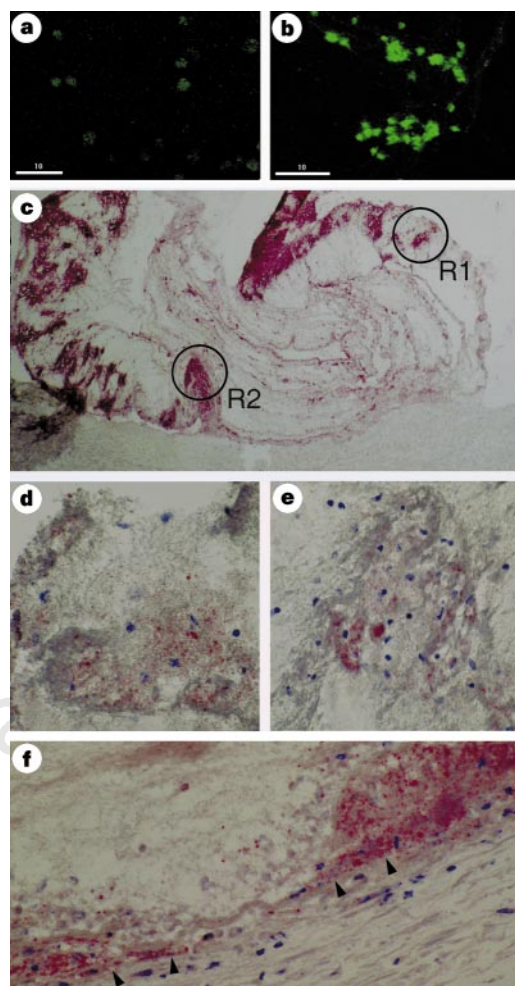


Figure 5 Expression of CD40L on activated platelets *in vivo*. With the method chosen, CD40L was specifically detected on the surface of activated platelets, as demonstrated by confocal microscopy of **a**, resting, and **b**, thrombin-activated platelets stained with mAb TRAP1 (shown are digitized images; bars, $10 \mu\text{m}$). **c**, Photomicrograph of a fresh intravascular thrombus, resulting from an emergency catheterization of the femoral vein. Staining for gplb α (CD42b) highlights laminar aggregates formed by platelets at the thrombus-vessel interface and across the thrombus (original magnification, $\times 10$). **d**, **e**, Immunohistological detection of CD40L with TRAP1 on aggregating platelets in characteristic regions (R1, R2) of the thrombus shown in **c** ($\times 66$; consecutive serial tissue sections). **f**, Expression of CD40L on aggregating platelets along the thrombus-vessel interface (arrows) of a thrombus developing on the surface of an atherosclerotic plaque ($\times 86$). The data are representative of five fresh intravascular thrombi examined.

avidin–phycoerythrin. HUVEC were analysed with monoclonal antibodies 1.2B6 (CD62E; Ancell), 51-10C9 (CD106; Pharmingen) or 6.5B5 (CD54; Dako).

Confocal microscopy and immunohistochemistry. Resting or thrombin-activated platelets were spread on slides, dried (1 h), fixed with acetone (10 min), rehydrated in detergent-free PBS, and stained with TRAP1 (50 µg ml⁻¹ in 3% heat-inactivated rat serum for 3 h). With washing steps in between, slides were incubated with biotinylated rat-anti-mouse antibodies (Jackson; 1:100 in 3% rat serum for 90 min), then streptavidin–TRITC (Jackson; 1:150) and analysed by confocal microscopy (Leica). Frozen tissue sections were fixed with acetone (10 min), rehydrated in PBS, and stained in detergent-free buffer with TRAP1 or gpIbα-specific antibody 142.11 (ref. 25) using the APAAP labelling technique²⁶ with new fuchsin as chromogen.

Immunoprecipitation and western blotting. CD40L was immunoprecipitated from platelets (1.5 × 10⁸) or activated CD4⁺ T cells (1.5 × 10⁶) using TRAP1, TRAP2 (ref. 8), or a chimaeric CD40–Ig reagent⁹, size-separated on SDS–PAGE (12% gel), blotted, and stained with anti-CD40L serum²⁷ using the Western Light Plus-kit (Tropix).

Amplification by PCR and sequencing. Platelets obtained by apheresis were filtered (PL50 leukocyte-removal filter; Pall, UK) and analysed by flow cytometry (contamination by leukocytes was estimated as <1 in 3,000). Total RNA from platelets and the megakaryoblastic lines MEG-01 (ref. 28) and UT-7 (ref. 29) was reverse transcribed into cDNA using oligo(dT), amplified by PCR with CD40L-specific primers²⁷ and sequenced.

Received 8 October; accepted 14 November 1997.

- Armitage, R. J. *et al.* Molecular and biological characterization of a murine ligand for CD40. *Nature* **357**, 80–82 (1992).
- Graf, D., Korthäuer, U., Mages, H. W., Senger, G. & Kroczeck, R. A. Cloning of TRAP, a ligand for CD40 on human T cells. *Eur. J. Immunol.* **22**, 3191–3194 (1992).
- Lederman, S. *et al.* Identification of a novel surface protein on activated CD4⁺ T cells that induces contact-dependent B cell differentiation (help). *J. Exp. Med.* **175**, 1091–1101 (1992).
- Gauchat, J. F. *et al.* Induction of human IgE synthesis in B cells by mast cells and basophils. *Nature* **365**, 340–343 (1993).
- Kroczeck, R. A. *et al.* Defective expression of CD40 ligand on T cells causes X-linked immunodeficiency with hyper-IgM (HIGM1). *Immunol. Rev.* **138**, 39–59 (1994).
- Nieuwenhuis, H. K., van Oosterhout, J. J., Rozenmuller, E., van Iwaarden, F. & Sixma, J. J. Studies with a monoclonal antibody against activated platelets: evidence that a secreted 53,000-molecular weight lysosome-like granule protein is exposed on the surface of activated platelets in the circulation. *Blood* **70**, 838–845 (1987).
- Holmsen, H. in *Hemostasis and Thrombosis* (eds Colman, R. W., Hirsh, J., Marder, V. J. & Salzman, E. W.) 3rd edn, 524–545 (Lippincott, Philadelphia, 1994).
- Graf, D. *et al.* A soluble form of TRAP (CD40 ligand) is rapidly released after T cell activation. *Eur. J. Immunol.* **25**, 1749–1754 (1995).
- Lauffer, L. *et al.* Monomeric and dimeric forms of soluble receptors can differ in their neutralization potential. *Behring Inst. Mitt.* **96**, 21–31 (1995).
- Yellin, M. J. *et al.* Functional interactions of T cells with endothelial cells: the role of CD40L–CD40-mediated signals. *J. Exp. Med.* **182**, 1857–1864 (1995).
- Hollenbaugh, D. *et al.* Expression of functional CD40 by vascular endothelial cells. *J. Exp. Med.* **182**, 33–40 (1995).
- Karmann, K., Hughes, C. C., Schechner, J., Fanslow, W. C. & Pober, J. S. CD40 on human endothelial cells: inducibility by cytokines and functional regulation of adhesion molecule expression. *Proc. Natl Acad. Sci. USA* **92**, 4342–4346 (1995).
- Springer, T. A. Traffic signals for lymphocyte recirculation and leukocyte emigration: the multistep paradigm. *Cell* **76**, 301–314 (1994).
- Bevilacqua, M. P. Endothelial-leukocyte adhesion molecules. *Annu. Rev. Immunol.* **11**, 767–804 (1993).
- Ebnet, K., Kaldjian, E. P., Anderson, A. O. & Shaw, S. Orchestrated information transfer underlying leukocyte endothelial interactions. *Annu. Rev. Immunol.* **14**, 155–177 (1996).
- Baggiolini, M., Dewald, B. & Moser, B. Interleukin-8 and related chemotactic cytokines—CXC and CC chemokines. *Adv. Immunol.* **55**, 97–178 (1994).
- Ware, J. A. & Heistad, D. D. Platelet-endothelium interactions. *N. Engl. J. Med.* **328**, 628–635 (1993).
- Clark, R. A. F. in *The Molecular and Cellular Biology of Wound Repair* (ed. Clark, R. A. F.) 2nd edn, 3–50 (Plenum, New York, 1996).
- Alderson, M. R. *et al.* CD40 expression by human monocytes: regulation by cytokines and activation of monocytes by the ligand for CD40. *J. Exp. Med.* **178**, 669–674 (1993).
- Caux, C. *et al.* Activation of human dendritic cells through CD40 cross-linking. *J. Exp. Med.* **180**, 1263–1272 (1994).
- Lauffer, P. A. *et al.* Stimulation of CD40 with purified soluble gp39 induces proinflammatory responses in human monocytes. *J. Immunol.* **155**, 4917–4925 (1995).
- Hawrylowicz, C. M., Howells, G. L. & Feldman, M. Platelet-derived interleukin 1 induces human endothelial adhesion molecule expression and cytokine production. *J. Exp. Med.* **174**, 785–790 (1991).
- Kaplanski, G. *et al.* Activated platelets induce endothelial secretion of interleukin-8 *in vitro* via an interleukin-1-mediated event. *Blood* **81**, 2492–2495 (1993).
- Davies, M. J. The contribution of thrombosis to the clinical expression of coronary atherosclerosis. *Thromb. Res.* **82**, 1–32 (1996).
- Kenny, D., Newman, P. J., Morattek, P. A. & Montgomery, R. A. A dinucleotide deletion results in defective membrane anchoring and circulating soluble glycoprotein Iba in a novel form of Bernard-Soulier syndrome. *Blood* **90**, 2626–2633 (1997).
- Cordell, J. L. *et al.* Immuno-enzymatic labeling of monoclonal antibodies using immune complexes of alkaline phosphatase and monoclonal anti-alkaline phosphatase (APAAP complexes). *J. Histochem. Cytochem.* **32**, 219–229 (1984).
- Korthäuer, U. *et al.* Defective expression of T-cell CD40 ligand causes X-linked immunodeficiency with hyper-IgM. *Nature* **361**, 539–541 (1993).

- Ogura, M. *et al.* Establishment of a novel human megakaryoblastic leukemia cell line, MEG-01, with positive Philadelphia chromosome. *Blood* **66**, 1384–1392 (1985).
- Komatsu, N. *et al.* Establishment and characterization of a human leukemic cell line with megakaryocytic features: dependency on granulocyte-macrophage colony-stimulating factor, interleukin 3, or erythropoietin for growth and survival. *Cancer Res.* **51**, 341–348 (1991).

Acknowledgements. J.R.S. is a postdoctoral fellow of the Max-Planck Society. We thank R. Kurlle for the generous gift of CD40-Ig, J. Schröder for ELISA reagents for determination of IL-8, and H.W. Mages for critically reading the manuscript.

Correspondence and requests for materials should be addressed to R.A.K. (e-mail: KroczeckR@rki.de).

An African HIV-1 sequence from 1959 and implications for the origin of the epidemic

Tuofu Zhu[†], Bette T. Korber[‡], Andre J. Nahmias^{||}, Edward Hooper[§], Paul M. Sharp[#] & David D. Ho^{*}

^{*} Aaron Diamond AIDS Research Center, The Rockefeller University, 455 First Avenue, New York, New York 10016, USA

[‡] Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

[§] The Santa Fe Institute, Santa Fe, New Mexico 87501, USA

^{||} Department of Pediatrics, Emory University School of Medicine, Atlanta, Georgia 30303, USA

[¶] PO Box 4087, Worthing BN14 7LQ, UK

[#] Department of Genetics, University of Nottingham, Queens Medical Centre, Nottingham NG7 2UH, UK

There is considerable genetic diversity among viruses of different subtypes (designated A to J) in the major group of human immunodeficiency virus type 1 (HIV-1), the form of HIV that is dominant in the global epidemic^{1–3}. If available, HIV-1 sequences pre-dating the recognition of AIDS could be crucial in defining the time of origin and the subsequent evolution of these viruses in humans. The oldest known case of HIV-1 infection was reported to be that of a sailor from Manchester who died of an AIDS-like illness in 1959 (refs 4–6); however, the authenticity of this case has not been confirmed^{7,8}. Genetic analysis of sequences from clinical materials obtained from 1971 to 1976 from members of a Norwegian family infected earlier than 1971 showed that they carried viruses of the HIV-1 outlier group^{9,10}, a variant form that is mainly restricted to West Africa¹. Here we report the amplification and characterization of viral sequences from a 1959 African plasma sample that was previously found to be HIV-1 seropositive¹¹. Multiple phylogenetic analyses not only authenticate this case as the oldest known HIV-1 infection, but also place its viral sequence near the ancestral node of subtypes B and D in the major group, indicating that these HIV-1 subtypes, and perhaps all major-group viruses, may have evolved from a single introduction into the African population not long before 1959.

A serological study of 1,213 plasma samples obtained from Africa between 1959 and 1982 has been reported¹¹. Twenty-one samples were initially found to be HIV-1 seropositive by immunoassay, but only one was confirmed as reactive with HIV-1 by immunofluorescence, western blotting and radioimmunoprecipitation methods. This positive plasma sample, designated L70, was obtained in early 1959 from an adult Bantu male, with a sickle-cell trait and a glucose-6-phosphate-dehydrogenase deficiency, living in Leopoldville, Belgian Congo (now Kinshasa, Democratic Republic of Congo)^{11,12}. The viral sequences contained in this sample might provide insights into the evolution of HIV-1.

Because of the limited amount of plasma available from this sample and uncertainty about its condition, efforts were made to increase the likelihood of recovering HIV-1 sequences by RT-PCR

[†] Present address: University of Washington, Health Sciences Building, Room T239X, 1959 NE Pacific Avenue, Seattle, Washington DC 98195, USA.

(reverse transcription followed by polymerase chain reaction). Multiple primers were used in a single RT reaction, and all synthesized complementary DNAs were amplified by PCR using primers designed to amplify HIV-1 sequences from all known subtypes (see Methods). However, attempts to amplify HIV-1 fragments of >300 base pairs (bp) were unsuccessful, probably because of a low level of intact viral RNA in this old sample. However, after numerous attempts, four shorter sequences were obtained. As shown in Fig. 1, fragments a, b and c corresponded to *env*-gene regions containing the V3 loop, the gp120–gp41 junction (including the Rev-responsive element, RRE), and the carboxy-terminal domain of gp41, respectively; fragment d corresponded to a region of the *pol* gene. Each PCR product was then purified, cloned and sequenced.

Nucleotide sequences from these HIV-1 fragments were designated ZR59a, ZR59b, ZR59c and ZR59d (GenBank accession numbers AF030526–AF030544, AF030637–AF030651, AF030652–AF030671 and AF030672–AF030686 for a, b, c and d, respectively). Sequences were aligned to available sequences in the 1996 Los Alamos Database using multiple aligned sequence editor¹³. Between 6 and 17 HIV-1 sequences were obtained from each region, and each set of ZR59 sequences formed a tight cluster with only 0–3% divergence when analysed phylogenetically using the neighbour-joining method. Each sequence was also run against those in GenBank using basic local alignment search tool¹⁴ to find the closest match. Maximum sequence identity scores of 92% (ZR59a), 96% (ZR59b) and 94% (ZR59c) confirmed that the ZR59 sequences

are indeed unique and unlikely to be the result of PCR contamination.

Consensus sequences for ZR59a, ZR59b and ZR59c were concatenated for in-depth studies. Phylogenetic analyses were performed in the laboratories of B.T.K. and P.M.S. using various approaches and slightly different alternative alignments, but with similar results and conclusions. Three methods were used for phylogenetic analysis of the ZR59a–c sequences, namely, minimum evolution (neighbour joining), maximum likelihood and weighted parsimony (see Methods); the results of all three approaches were similar. The ZR59 sequence was positioned close to the ancestral node of subtypes B, D and F, although with a slightly preferred association with the D lineage (Fig. 2). Although the ZR59d sequence was phylogenetically less informative (because of a higher degree of sequence conservation in *pol*), analyses of this fragment resulted in a similar conclusion.

Several HIV-1 isolates are hybrids of different subtypes¹⁵, and the phylogenetic position found for ZR59 could be an artefact of it being a recombinant containing parts of modern subtypes D and B or F. Therefore, the ZR59 sequences were scanned using RIP¹⁶, a program that looks for the possibility of recombination in a query sequence relative to a set of control sequences. In this case, ZR59 was compared with the consensus sequence of each subtype within the HIV-1 major group. In a similarity comparison, all four ZR59 fragments were slightly closer to the B-clade consensus, but in the *env* fragments there were short stretches that were more similar to fragments of the D clade. However, overall, there was no statistically

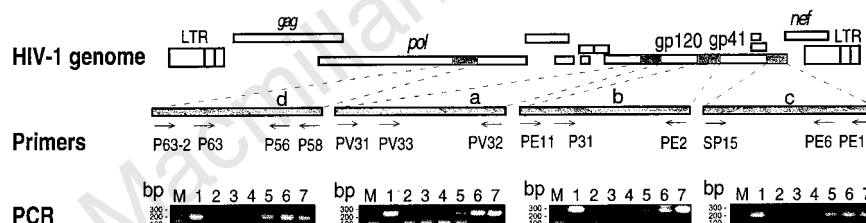


Figure 1 Detection of HIV-1 sequences in the L70 plasma sample by RT-PCR. The HIV-1 genome is shown at the top. ZR59 fragments (a, b, c and d) that were successfully amplified are shown in the middle, as are the primers used. Results of the successful RT-PCR are shown on the bottom. M, molecular marker; lane 1, positive plasma control; lane 2, negative plasma control; lane 3, reagent control

(sterile water); lane 4, L70 plasma without reverse transcription before PCR; lane 5, L70 plasma; lane 6, serum from an infected patient identified in 1985; lane 7, serum from an infected patient (H.) identified in 1994. Each gel corresponds to the ZR59 fragment it is shown under. LTR, long terminal repeat.

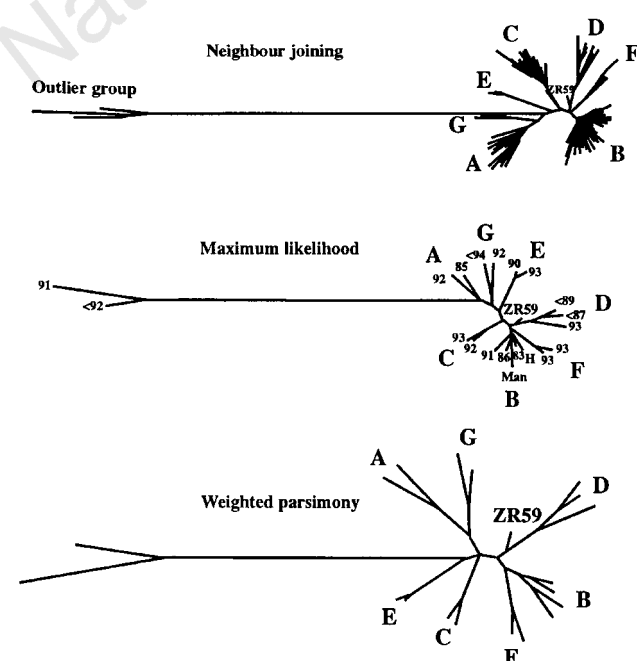


Figure 2 Phylogenetic analyses of the ZR59 sequence. ZR59 branched off the D clade near the B/D/F root in all analyses. The taxa are labelled with the year of sampling of a given sequence, or with '<' and the year of the primary publication if the year of sampling was not specified. The analyses included the following reference sequences and their GenBank accession numbers: A 92, 92RW020 (U08794); A 85, U455 (M62320); B 91, 91HT652 (U08443); B 83, LAI (X01762); B 86, JRFL (U63632); H, the internal control for this study sampled in 1994; Man, the Manchester sailor sequence⁷ (U23487); C 92, 92BR025 (U09132); C 93, 93MW965 (U08455); D <89, NDK (M27323); D <87, Z6 (K03458); D 93, 92ZR001 (U27419); E 90, CM240 (U54771); E 93, 93TH966 (U08456); F 93, 93BR020 (U27401); F 93, 93BR029 (U27413); G <94, LBV217 (U09664); G 92, 92UG975 (U22010); outlier group 91, VAU (X80020); and outlier group <92, MVP5180 (L20571). The same set of input taxa was used for the weighted-parsimony and maximum-likelihood trees, except that H was included in the maximum-likelihood tree only. A larger set of taxa was used for the neighbour-joining tree. The scale for branch lengths is comparable for the maximum-likelihood and neighbour-joining trees. For the weighted-parsimony tree, the branch lengths are not directly comparable to the branch lengths of the other two trees, but the branching pattern substantiates the pattern obtained by the other two methods.

significant evidence, using RIP¹⁶ or other approaches¹⁵, for a recombination crossover point. Thus, ZR59 is not a mosaic of modern sequences. If ZR59 were a result of a recombination of the B and D subtypes soon after they diverged, the recombination would be of little consequence and would be unlikely to be detectable, especially with the short sequences available.

For most regions of the HIV-1 genome, subtypes B and D are more closely associated with each other than are any other subtypes within the major group¹⁻³. However, phylogenetic analysis of the *env* regions showed an unusual association between viruses of subtypes B and F. RIP analysis of the entire *env* gene showed that subtype B is more similar to subtype F than to subtype D, specifically over two short stretches, encompassing the V3 and RRE regions, that coincide with fragments a and b. This finding explains the unusual B/D/F clustering found in our phylogenetic analyses (Fig. 2).

The rate of evolution of HIV-1 sequences has been estimated^{17,18}, although a precise translation of genetic distances into years is not possible for a single virus because of variation in the rate of evolution among different individuals or different lineages^{19,20}. Nevertheless, because of the high evolutionary rate of HIV-1, a virus from 1959 should have evolved to a substantially lesser extent than viruses isolated in the 1980s or later. To test this hypothesis, a comparison was made between the likelihood of the original tree and that of a tree that had an elongated branch between the ZR59 sequence and the D ancestral node, with an imposed branch length comparable to that found in more modern isolates. The branch length of a contemporary subtype D virus sequence to the D ancestral node was found for the sequence NDK, which was then used to fix the branch length for ZR59 to the ancestral node. The tree was subsequently optimized with respect to the new branch length using maximum likelihood, and the log likelihood values of the original tree and the modified tree were compared using the method of Kishino and Hasegawa²¹ as implemented by PHYLIP. The difference of 14.85 in log likelihood values for the two trees showed that the modified tree, with a long branch length for ZR59, was significantly less likely ($P < 0.05$). In other words, ZR59 displayed properties of an older sequence. A likelihood-ratio test was also used to test the placement of ZR59 as an early branch-off from the B clade rather than from the D clade, but there was no significant difference between the two alternatives. This conclusion is consistent with results from bootstrap analyses. Although the maximum-likelihood, weighted-parsimony and neighbour-joining trees of Fig. 2 yielded relatively high bootstrap values (83, 75 and 40, respectively), supporting the existence of a branch point that includes the B, D and F subtypes and ZR59, the association between ZR59 and clade D was weaker (bootstrap values of 56, <50 and 38 were obtained from the maximum-likelihood, weighted-parsimony and neighbour-joining trees, respectively). Thus, although a branching-off of ZR59 from the D lineage is the probable result from each of the tree-building methods, this conclusion is not definitive.

We estimated the branch length between ZR59 and the ancestral node of subtypes B, D and F by several methods. Using maximum likelihood, a distance (substitutions per site) of 0.023 was obtained with the fastDNaml and DNArates programs (Fig. 2). A distance of 0.037 was obtained with PAML using a REV model and optimized gamma distribution (see Methods and ref. 22). PHYLIP 3.6 was also used to reconstruct an ancestral sequence based on a maximum-likelihood estimate of the most probable character at each site in the nodal sequence. A distance of only 0.026 separated ZR59 from this reconstructed sequence. Thus, ZR59 is similar to the ancestral sequence of HIV-1 subtypes B, D and F. But are these sequences identical? The maximum-likelihood tree in Fig. 2 was compared with one in which ZR59 has a zero branch length from the B, D and F ancestral sequence. The latter tree was significantly less likely ($P < 0.05$; ref. 21) with a difference in log likelihood of 34.17,

indicating that ZR59 was probably slightly more recent than the B/D/F ancestor.

The short but non-zero distance from the common ancestor of subtypes B and D (and F) to ZR59 indicates that ancestral HIV-1 of subtypes B and D (and F) must have existed before 1959, although probably only a few years before then. This is about a decade earlier than the previous estimate for the divergence of B and D clades¹⁷. This finding also refutes the suggestion that HIV-1 subtype-B infection was responsible for AIDS-like syndromes beginning in the 1930s in various European populations²³. Our results also indicate that subtypes B, D and F may have evolved with the human population rather than arising from multiple cross-species transmission events^{1,22}.

The phylogenetic analyses (Fig. 2) show that ZR59 is not too distant from the internal node for all major-group viruses; however, evolutionary rates may differ in different regions of the phylogenetic tree. Nonetheless, both the major- and the outlier-group viruses were clearly present in humans by 1960 (ref. 10). Our results, the rate of HIV-1 evolution (around 0.005–0.01 nucleotide changes per site per year¹⁷) and previously described methods of estimation of evolutionary rates²⁴ indicate that the major-group viruses that dominate the global AIDS pandemic at present shared a common ancestor in the 1940s or the early 1950s. Given their 'starburst' phylogeny, HIV-1 was probably introduced into humans shortly before that time frame, about a decade or two earlier than previously estimated^{17,25}. Thus, given the large genetic distance between HIV-1 and HIV-2, the divergence of these viruses could not have occurred in the late 1940s (ref. 25); that branching point must have been considerably earlier^{17,26–28}. The diversification of HIV-1 in the past 40–50 years portends even greater viral heterogeneity in the coming decades, and underscores the need for continued surveillance. The factors that propelled the initial spread of HIV-1 in central Africa remain unknown: the role of large-scale vaccination campaigns, perhaps with multiple uses of non-sterilized needles, should be carefully examined, although social changes such as easier access to transportation, increasing population density and more frequent sexual contacts may have been more important. □

Methods

RNA isolation and reverse transcription. The L70 plasma sample (200 µl) was diluted with 10 ml of precooled (4 °C) PBS and centrifuged at 40,000 r.p.m. for 4 h at 4 °C in a swing rotor (Sorvall SW40 TI). The pellet was then resuspended in 140 µl of precooled (4 °C) PBS. A QIAamp HCV kit (QIAGEN) was used to extract the viral RNA, which was eluted in 50 µl of RNase-free sterile water. The viral RNA was incubated with RNase-free DNase in the presence of 15 mM MgCl₂ for 20 min at 37 °C, and then incubated for 5 min at 80 °C to end the reaction. The RNA (25 µl) was mixed with 40 ng of each of the following antisense primers: P58 (residues 3798–3776, according to the HXB2 sequence in the Los Alamos Database; sequence 5'-GAC AAA CTC CCA CTC AGG AAT CCA-3'), PV32 (residues 7258–7235, 5'-TAC CTG TTG TAA AGT GTT ATT CCA-3'), PE2 (residues 7956–7934, 5'-GCC TGG AGC TGT TTA ATG CCC CA-3'), and PE12 (residues 8871–8844, 5'-CTG GCT CAG CTC GTC TCA TTC TTT CCC T-3'). This mixture was heated to 70 °C for 10 min and used to synthesize HIV-1 cDNA at 42 °C for 50 min in a solution containing 50 mM KCl, 10 mM Tris-HCl (pH 9.0), 0.1% Triton X-100, 0.3 mM of each of the deoxynucleoside triphosphates (dNTPs), 2.5 mM MgCl₂ and 40 units AMV reverse transcriptase (Promega). The mixture was finally heated at 70 °C for 15 min to inactivate the reverse transcriptase.

PCR and DNA sequencing. All newly synthesized cDNA was used simultaneously to amplify multiple HIV-1 sequences with primers P58 and P63 (residues 3580–3603, 5'-GCC ATT TAA AAA TCT GAA AAC AGG-3') for the *pol* region, PV32 and PV31 (residues 7019–7050, 5'-GCA GAA GAA GAG GTA GTA ATT AGA TCT GAA AA-3') for the V3 region, PE2 and PE11 (residues 7646–7671, 5'-ATG AGG GAC AAT TGG AGA AGT GAA TT-3') for the gp120/gp41-junction region, and PE12 and SP15 (residues 8600–8622, 5'-CTC AAA TAT TGG TGG AAT CTC CT-4') for the C terminus of gp41 (Fig. 1). The reaction buffer was the same as that described for reverse transcription,

except that the reverse transcriptase was replaced by four units of *Thermus aquaticus* DNA polymerase (Promega). Amplification cycles for the first round of the PCR were 96 °C for 4 min, 94 °C for 50 s, 55 °C for 30 s, and 72 °C for 1 min for 32 cycles, followed by a final extension at 72 °C for 10 min. The first-round PCR products (4 µl) were used in a second-round PCR to amplify specific HIV-1 regions with primer pairs P63 and P56 (residues 3757–3734, 5'-TGT CCA CCA TGC TTC CCA TGT TTC-3') for the *pol* region, PV32 and PV33 (residues 7053–7079, 5'-TCA CAG ACA ATG CTA AAA CCA TAA TAG-3') for the V3 region, PE2 and P31 (residues 7704–7736, 5'-TAG GAG TAG CAC CCA CCA AGG CAA AGA GAA GAG-3') for the gp120–gp41-junction region, or SP15 and PE6 (residues 8822–8796, 5'-ACT ACT TTT TGA CCA CTT GCC ACC CAT-3') for the C terminus of gp41. Amplification conditions were 94 °C for 2 min, 94 °C for 30 s, 55 °C for 25 s, and 72 °C for 1 min for 35 cycles, followed by a final extension at 72 °C for 10 min. All PCR reactions were carried out in the Perkin–Elmer model 9600 thermocycler. PCR products were purified with the QIAEX II gel extraction kit (QIAGEN) and then inserted into the PCR3.1 vector (Invitrogen) before DNA sequencing in an automated sequencer (ABI PRISM 377).

Phylogenetic analyses. The neighbour-joining tree was generated using PHYLIP 3.5 (J. Felsenstein). We used the 111 sequences from the *env* alignment (in the Los Alamos Database) that spanned all three regions of interest, a total of 468 bases in the alignment after gap-stripping. By using all available taxa, we could determine that the phylogenetic behaviour of the ZR59 sequence was unique, and that no sequences from the 1980s or 1990s gave a similar result. The distance matrix for the neighbour-joining tree was created using the F84 option of the program DNADIST with the likelihood estimates of the relative rates of site mutations described below (when a simpler Kimura two-parameter model was used to calculate the distance matrix, the ZR59 sequence preferentially branched off from the B clade in the neighbour-joining tree). We selected a representative subset of the 111 sequences to represent the major-group viruses; we tried to include sequences with known years of sampling. These sequences and the ZR59 sequence were subjected to tree-building methods that are computationally intensive, namely maximum likelihood and weighted parsimony. Maximum-likelihood analyses were done using several programs. FastDNAm1²⁹ was used for the bootstrap analysis³⁰ and the initial maximum-likelihood tree. PHYLIP 3.5 was used for likelihood-ratio tests to compare the trees. We used PHYLIP 3.6 (J. Felsenstein) to reconstruct the most probable sequence at the ancestral nodes. All likelihood trees incorporated an estimate of rate variation between sites, an important element of accurate tree reconstruction, using the DNARates program (G. Olsen, personal communication) which optimizes the relative rate of substitution at each position in an alignment using maximum likelihood. A maximum-likelihood tree incorporating a REV model with an optimized gamma distribution using PAML (version 1.1, Z. Yang) was also generated²²; this tree confirmed results obtained with fastDNAm1. Statistical comparisons of the maximal-likelihood trees were made using the method of Kishino and Hasegawa²¹. When using weighted parsimony, we estimated substitution rates between the four bases from a parsimony tree generated by PAUP (D. Swofford), using the MacClade program; we then used the inverse of the observed substitution frequency to weight the character changes, by PAUP, in a subsequent parsimony tree. The accuracy of the branching pattern can be improved markedly by this approach, but the branch lengths do not reflect true distances.

Received 18 September; accepted 17 December 1997.

1. Sharp, P. M., Robertson, D. L., Gao, F. & Hahn, B. H. Origins and diversity of human immunodeficiency viruses. *AIDS* **8**, S27–S43 (1994).
2. Gao, F. *et al.* Molecular cloning and analysis of functional envelope genes from human immunodeficiency virus type 1 sequence subtypes A through G. *J. Virol.* **70**, 1651–1667 (1996).
3. Louwagie, J. *et al.* Phylogenetic analysis of gag genes from 70 international HIV-1 isolates provides evidence for multiple genotypes. *AIDS* **7**, 769–780 (1993).
4. Williams, G., Stretton, T. B. & Leonard, J. C. Cytomegalic inclusion disease and pneumocystis carinii infection in an adult. *Lancet* **ii**, 951–955 (1960).
5. Williams, G., Stretton, T. B. & Leonard, J. C. AIDS in 1959? *Lancet* **ii**, 1136 (1983).
6. Corbett, G., Bailey, A. S. & Williams, G. HIV infection in Manchester, 1959. *Lancet* **336**, 51 (1990).
7. Zhu, T. & Ho, D. D. Was HIV present in 1959? *Nature* **374**, 503–504 (1995).
8. Hooper, E. & Hamilton, W. D. 1959 Manchester case of syndrome resembling AIDS. *Lancet* **348**, 1363–1365 (1996).
9. Froland, S. S. *et al.* HIV-1 infection in Norwegian family before 1970. *Lancet* **i**, 1344–1345 (1988).
10. Jonassen, T. O. *et al.* Sequence analysis of HIV-1 group O from Norwegian patients infected in the 1960s. *Virology* **231**, 43–47 (1997).
11. Nahmias, A. J. *et al.* Evidence for human infection with an HTLV-III/LAV-like virus in central Africa, 1959. *Lancet* **i**, 1279–1280 (1986).
12. Motulsky, A. G. *et al.* Population genetic studies in the Congo. 1. Glucose-6-phosphate dehydrogenase

- deficiency, hemoglobin S, and malaria. *Am. J. Hum. Genet.* **18**, 514–537 (1966).
13. Faulkner, D. V. & Jurka, A. Multiple aligned sequence editor (MASE). *Trends Biochem. Sci.* **13**, 321–322 (1988).
14. Altschul, S. F., Gish, W., Miller, W. & Myers, E. W. Basic local alignment search tool (BLAST). *J. Mol. Biol.* **215**, 403–410 (1990).
15. Robertson, D. L., Sharp, P. M., McCutchan, F. E. & Hahn, B. H. Recombination in HIV-1. *Nature* **374**, 124–126 (1995).
16. Siepel, A. C., Halpern, A. L., Macken, C. & Korber, B. A computer program designed to rapidly screen for HIV-1 intersubtype recombinant sequences. *AIDS Res. Hum. Retroviruses* **11**, 1413–1416 (1995).
17. Li, W.-H., Tanimura, M. & Sharp, P. M. Rates and dates of divergence between AIDS-virus nucleotide sequences. *Mol. Biol. Evol.* **5**, 313–330 (1988).
18. Kelly, J. K. An application of population genetic theory to synonymous gene sequence evolution in the human immunodeficiency virus (HIV). *Genet. Res.* **64**, 1–9 (1994).
19. Lukashov, V. V., Kuiken, C. L. & Goudsmit, J. Intrahost human immunodeficiency virus type 1 evolution is related to length of the immunocompetent period. *J. Virol.* **69**, 6911–6916 (1995).
20. Wolinsky, S. *et al.* Adaptive evolution of human immunodeficiency virus type 1 during the natural course of infection. *Science* **277**, 537–542 (1996).
21. Kishino, H. & Hasegawa, M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.* **29**, 170–179 (1989).
22. Leitner, T., Escanilla, D., Franzen, C., Uhlen, M. & Albert, J. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proc. Natl Acad. Sci. USA* **93**, 10864–10869 (1996).
23. Goudsmit, J. *Viral Sex: the Nature of AIDS*. (Oxford Univ. Press, 1997).
24. Leitner, T. S., Kumar, S. & Albert, J. Tempo and mode of nucleotide substitutions in gag and env gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. *J. Virol.* **71**, 4761–4770 (1997).
25. Smith, T. F., Srinivasan, A., Schochetman, G., Marcus, M. & Myers, G. The phylogenetic history of immunodeficiency viruses. *Nature* **333**, 573–575 (1988).
26. Sharp, P. M. & Li, W.-H. Understanding the origins of AIDS viruses. *Nature* **336**, 315 (1988).
27. Eigen, M. & Nieselt-Struwe, K. How old is the immunodeficiency virus? *AIDS* **4** (suppl.) S85–S93 (1990).
28. Gojbori, T. *et al.* Evolutionary origin of human and simian immunodeficiency viruses. *Proc. Natl Acad. Sci. USA* **87**, 4108–4111 (1990).
29. Olsen, G. J., Matsuda, H., Hagstrom, R. & Overbeek, R. fastDNAm1: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput. Appl. Biosci.* **10**, 41–48 (1994).
30. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376 (1981).

Acknowledgements. We thank J. Felsenstein for an early release of PHYLIP 3.6 and for helpful suggestions, T. Leitner for advice, and G. Olsen for providing DNA rates and a preprint describing a method for estimating the relative rates of variation at each site. This study was supported by grants from the NIH, the Magic Johnson Foundation and the Irene Diamond Fund, and by donations from D. Pels, A. Gund and V. D. Adler.

Correspondence and requests for materials should be addressed to D.D.H. (e-mail: dho@adarc.org).

Retinoblastoma protein recruits histone deacetylase to repress transcription

Alexander Brehm*, Eric A. Miska*, Dennis J. McCance*†, Juliet L. Reid*, Andrew J. Bannister* & Tony Kouzarides*

* Wellcome/CRC Institute and Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QR, UK

† Department of Microbiology and Immunology, University of Rochester, Rochester, New York 14642, USA

The retinoblastoma protein (Rb) silences specific genes that are active in the S phase of the cell cycle and which are regulated by E2F transcription factors¹. Rb binds to the activation domain of E2F and then actively represses the promoter by a mechanism that is poorly understood^{2,3}. Here we show that Rb associates with a histone deacetylase, HDAC1, through the Rb 'pocket' domain. Association with the deacetylase is reduced by naturally occurring mutations in the pocket and by binding of the human papilloma virus oncoprotein E7. We find that Rb can recruit histone deacetylase to E2F and that Rb cooperates with HDAC1 to repress the E2F-regulated promoter of the gene encoding the cell-cycle protein cyclin E. Inhibition of histone deacetylase activity by trichostatin A (TSA) inhibits Rb-mediated repression of a chromosomally integrated E2F-regulated promoter. Our results indicate that histone deacetylases are important for regulating the cell cycle and that active transcriptional repression by Rb may involve the modification of chromatin structure.

Histone deacetylase activity has been ascribed to two mammalian proteins, HDAC1 and HDAC2 (refs 4, 5). These deacetylases exist in

a complex with other proteins and are implicated in transcriptional repression resulting from nucleosome remodelling^{6–12}. Given that a known Rb-binding protein, Rbp48 (refs 13, 14), is found in the deacetylase complex, we investigated whether the Rb repressor is associated with histone deacetylase activity. Figure 1a shows that a fusion protein of Rb with glutathione-S-transferase (GST) can purify histone deacetylase activity from nuclear extracts derived from HeLa cells, whereas several GST fusions to other nuclear proteins fail to bind significant amounts of deacetylase. Association with deacetylase activity is also seen in two other cell types (U2OS and lymphocytes transformed by Epstein–Barr virus (EBV)) and the activity is sensitive to nanomolar concentrations of the histone deacetylase inhibitor TSA (data not shown). Furthermore, Rb-specific antibodies precipitate deacetylase activity from extracts of mammalian cells (Fig. 1b). These results indicate that Rb is associated with deacetylase activity *in vivo*.

The tumour-suppressor functions of Rb reside in the 'pocket' domain. We find that the integrity of the pocket is crucial for binding deacetylase (Fig. 1c). A tumour-derived point mutation (C706F) or deletion (Δ ex21) abolishes binding of deacetylase, as do two carboxy-terminal truncations that impinge on or delete the pocket domain B.

Viral oncoproteins can stimulate E2F-regulated promoters by sequestering Rb¹. Binding of these viral proteins to the pocket domain is mediated by an LXCXE sequence motif. Given that histone deacetylase associates with the pocket, we investigated whether binding of the human papilloma virus (HPV) E7 oncoprotein could displace the deacetylase. Figure 1d shows that association of Rb with deacetylase is abrogated in the presence of E7. An E7 protein with a mutated, non-functional LXCXE motif (C24) fails to displace deacetylase efficiently compared with a mutation elsewhere in the protein (Δ 5–10). In addition, a peptide containing the Rb-binding LXCXE motif sequence of E7 is sufficient to displace

deacetylase activity from Rb. In contrast, a control peptide (haemagglutinin, HA) has no effect on deacetylase binding. These results suggest that E7 can effectively compete with histone deacetylase for occupancy of the Rb pocket. Our finding that GST–Rb can associate with deacetylase in nuclear extracts of HPV-transformed HeLa cells indicates that there is not enough E7 in these extracts to disrupt the interaction of histone deacetylase with excess GST–Rb.

Figure 2a shows that GST–Rb can complex with HDAC1 translated *in vitro* (lane 3). GST–Rb can also associate with HDAC1 and HDAC2 present in nuclear extract, as revealed by western blotting using anti-HDAC1/2 antibody (Fig. 2b). When Rb and epitope-tagged HDAC1 (HDAC1-Flag) are coexpressed by transient transfection of mammalian cells, an interaction is detected by western blot analysis following precipitation with either anti-Flag (Fig. 2c) or anti-Rb antibody (Fig. 2d). The interaction between Rb and HDAC1 can also be demonstrated in untransfected cells (Fig. 2e). In this experiment, an immunoprecipitate obtained with Rb-specific antibody was shown to contain HDAC1 (Fig. 2e, lane 3), whereas an immunoprecipitate obtained with an irrelevant antibody (anti-HA) did not contain HDAC1 (Fig. 2e, lane 2). Thus Rb associates with the HDAC1 deacetylase *in vivo*, consistent with the ability of Rb to complex with deacetylase activity in nuclear extracts (Fig. 1).

The integrity of the Rb pocket is important for interaction with HDAC1. Binding of *in vitro* translated HDAC1 to GST–Rb is compromised by the tumour-derived mutations C706F and Δ ex21 and by other pocket deletions (Fig. 2f). These Rb pocket mutants have also lost the ability to purify deacetylase activity from nuclear extracts (Fig. 1c). Consistent with HDAC1 binding to the pocket of Rb is the observation that the LXCXE peptide can displace HDAC1 from Rb (data not shown).

If Rb is able to recruit histone deacetylase to E2F, then E2F should be complexed with HDAC1 in the presence of Rb. Figure 3a shows that the activation domain of E2F can interact with *in vitro*

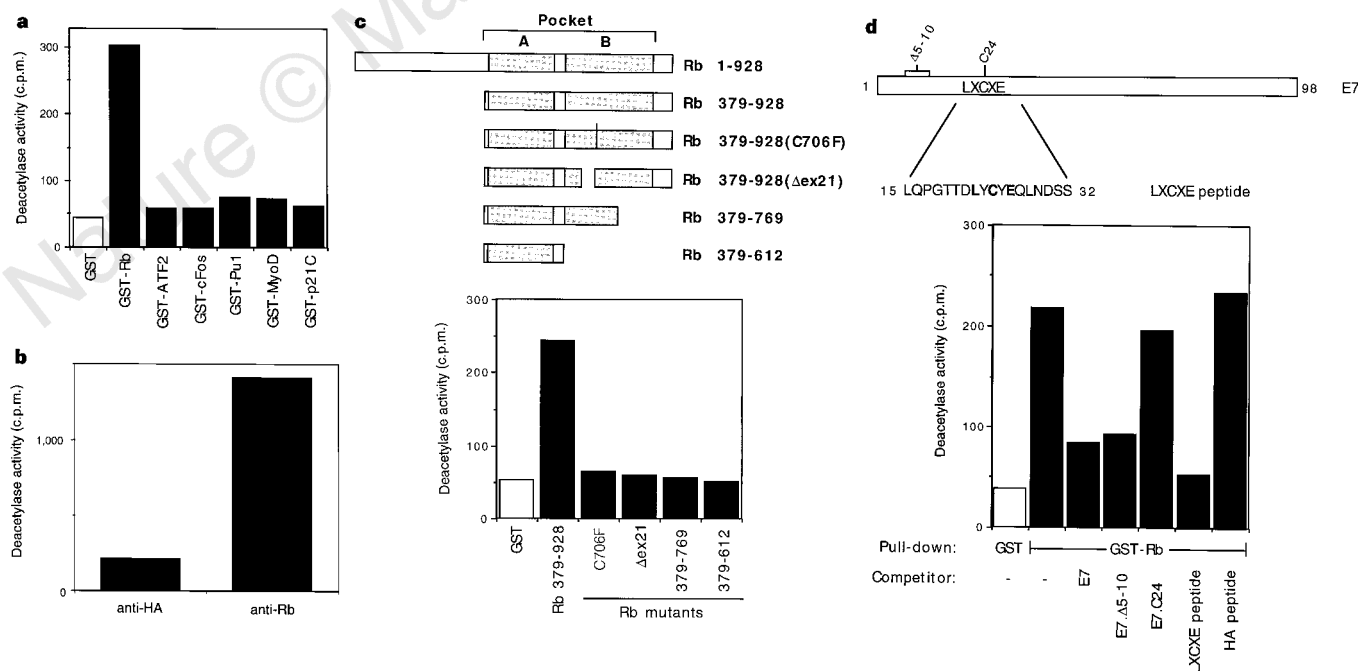


Figure 1 Rb associates with histone deacetylase activity *in vitro* and *in vivo*. **a**, GST–Rb fusion protein binds histone deacetylase activity from nuclear extract. Histone deacetylase activity is given as c.p.m. of ³H-acetate released from an acetylated histone H4 peptide⁴. **b**, Histone deacetylase activity coimmunoprecipitates with Rb. Jurkat nuclear extract was immunoprecipitated with 12C5A (anti-HA) or G3-245 (anti-Rb) antibody, as indicated. **c**, Binding of deacetylase activity is abrogated by pocket domain mutations. Top: representation of mutants. Pocket

subdomains A and B are represented as grey boxes. Bottom: binding of deacetylase activity by GST–Rb379–928 and pocket-domain mutants. **d**, HPV16 E7 competes Rb-associated deacetylase activity. LXCXE: Rb-binding motif. Deacetylase activity was bound in the presence and absence of purified recombinant E7 proteins (E7, E7 Δ 5–10, E7C24) or competitor peptides (HA, LXCXE; 20 μ g ml^{–1}), as indicated.

translated HDAC1 (lane 4). Several observations indicate that this interaction is indirectly mediated by the binding of Rb to E2F. First, mutating the Rb-binding site¹⁵ in E2F (E2FΔRb; Fig. 3a, lane 6) abolishes the interaction with HDAC1, whereas mutating the Mdm2-binding site¹⁶ (E2FΔMDM2; Fig. 3a, lane 5) has no effect. Second, addition of recombinant histidine-tagged Rb to the pull-down reaction (see Methods) increases the interaction of GST–E2F with HDAC1 considerably (Fig. 3b: compare lanes 1 and 2), whereas addition of a control protein, XRCC4, has no effect. These experiments suggest that the interaction between E2F and HDAC1 in reticulocyte lysate is indirectly mediated by the limiting amount of Rb present in the lysate. Coimmunoprecipitation shows that E2F can complex with HDAC1 in transfected cells (data not shown).

To confirm that HDAC1 recruited by Rb to E2F still retained its biological activity, we tested whether an E2F–Rb complex could associate with histone deacetylase. Figure 3c shows that a GST–E2F column does not purify deacetylase activity from nuclear extract (lane 2) unless it is first saturated with recombinant Rb (lane 5). Binding of deacetylase activity is dependent on Rb because GST–E2F bearing a mutated Rb-binding site (GST–E2FΔRb) fails to bind deacetylase activity in this assay (Fig. 3c, lane 6). These results indicate that Rb can recruit active HDAC1 enzyme to the E2F activation domain.

The finding that Rb can recruit HDAC1 to E2F raises the possibility that HDAC1 might function as a co-repressor of E2F. We tested this using the cyclin E promoter, an E2F-regulated target^{17,18}, and found that expression of HDAC1 moderately repressed cyclin E promoter activity (Fig. 4a). In the presence of coexpressed Rb, HDAC1 represses more efficiently and in a cooperative manner.

Given that Rb can tether HDAC1 to E2F, we investigated whether

Rb uses deacetylase activity to mediate repression of E2F. If so, then Rb-mediated repression should be sensitive to the histone deacetylase inhibitor TSA. We therefore used a cell line containing an integrated chloramphenicol acetyltransferase (CAT) gene reporter linked to Gal4-binding sites (3T3/208) (ref. 19). This reporter should be arranged as a complete chromatin structure. Figure 4b shows that the activity of this gene is stimulated by expression of a Gal4-E2F fusion (lane 2), and that coexpression of Rb results in promoter silencing (lanes 3 and 4), but that Rb-mediated repression is compromised in the presence of TSA (Fig. 4b: compare lanes 3 and 4 to lanes 7 and 8). The abrogation of Rb repression by TSA is also observed with non-integrated reporters under conditions in which a Gal4-dependent reporter is introduced by transient transfection (data not shown). These results confirm that Rb uses deacetylase activity to repress E2F.

Our results suggest a new mechanism for repression of E2F-regulated genes by Rb: Rb silences these genes by recruiting a repressive histone deacetylase to E2F (Fig. 4c). Once bound to the promoter, the deacetylase may convert the surrounding chromatin from a transcriptionally active (hyperacetylated) to a transcriptionally repressed (hypoacetylated) state. The demonstration that transfected DNA templates, which may not adopt a complete chromatin structure, are still responsive to HDAC1 repression and TSA-mediated stimulation^{5,9–11} (Fig. 4b, and data not shown) raises the possibility that deacetylases may target non-histone proteins involved in transcriptional regulation. This view is supported by the fact that the histone acetyl transferase CBP/p300 (refs 20, 21) regulates by acetylation the function of non-histone transcription factors²².

Rb binds to HDAC1 through its pocket domain using sequences that are distinct from the E2F-binding site. We have no evidence that

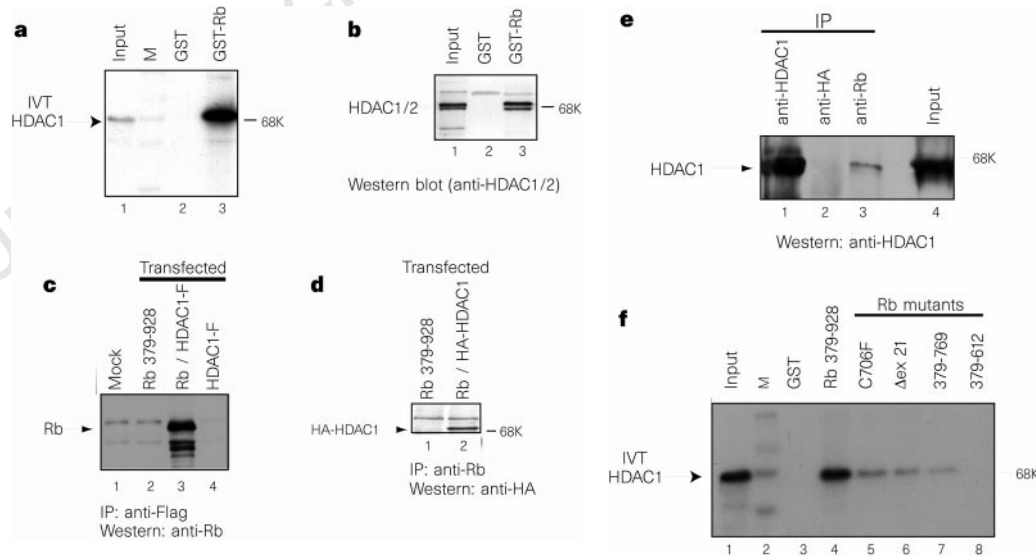


Figure 2 Rb interacts with histone deacetylases *in vitro* and *in vivo*. **a**, A GST–Rb fusion protein interacts with *in vitro*-translated ³⁵S-methionine-labelled HDAC1 in a GST pull-down assay. Lane 1: ³⁵S-HDAC1 input (2.5%). M: relative molecular mass marker. **b**, GST–Rb affinity-purifies histone deacetylases HDAC1 and HDAC2 from HeLa nuclear extract. Histone deacetylases were visualized by western blot analysis using the 1121 antibody, which recognizes both HDAC1 and HDAC2 (ref. 4). Lane 1: HeLa nuclear extract input (5%). **c**, Rb coimmunoprecipitates with Flag-tagged HDAC1 from extracts of transfected U2OS cells. Cells were transfected with pCMV-Rb379–928 and pcDNA3-HDAC1-F as indicated. Whole-cell extracts were precipitated with M2 (anti-Flag) antibody and the presence of Rb379–928 in immunoprecipitates was visualized by western blot analysis using XZ55 (anti-Rb) antibody. **d**, HA-tagged HDAC1 coimmunoprecipitates with Rb from extracts of transfected U2OS cells. Cells were transfected with pCMV-

Rb379–928 and pCMV-HA-HDAC1 as indicated. Whole-cell extracts were precipitated with XZ55 (anti-Rb) antibody and the presence of HA-HDAC1 in the immunoprecipitates was visualized by western blot analysis using 12C5A antibody. **e**, HDAC1 coimmunoprecipitates with Rb from Jurkat-cell extracts. Whole-cell extracts of untransfected Jurkat cells were precipitated with 1123 (anti-HDAC1; lane 1) (ref. 19), 12C5A (anti-HA; lane 2) and G3-245 (anti-Rb; lane 3) antibody, respectively. The presence of HDAC1 in immunoprecipitates was visualized by western blot analysis using 1121 (anti-HDAC1) antibody. Lane 4: Jurkat whole-cell extract input (1%). **f**, Mutations of the pocket domain of Rb compromise the interaction with HDAC1 *in vitro*. Binding of GST–Rb fusion proteins to *in vitro*-translated HDAC1 was assessed in a GST pull-down assay as indicated. Lane 1: ³⁵S-HDAC1 input (10%).

the interaction between Rb and HDAC1 is direct. The fact that Rb and HDAC1 do not interact in a yeast two-hybrid assay (data not shown) argues that the interaction in mammalian cells may be mediated by a component of the Sin3 complex—for example, by Rbp48 or by Sin3 itself^{6–12}.

The recruitment of histone deacetylase activity may explain why Rb has an active repression capacity when tethered to the promoter by a heterologous DNA-binding domain³. The ability to deliver deacetylase activity may allow Rb to silence any promoter to which it is recruited. Thus, the repression of pol I (ref. 23) and pol III genes^{24,25} by Rb may also involve the recruitment of deacetylase. Taken together, our results suggest that histone deacetylases are important in cell-cycle control. S-phase-specific genes, as well as other genes repressed by Rb, may be regulated, at least in part, at the chromatin level. □

Methods

Histone deacetylase assay. Histone deacetylase was assayed as described⁴.

Affinity purification and immunoprecipitation of histone deacetylase activity from cell extracts. Equivalent amounts of GST fusion proteins prebound to glutathione beads were added to 30 μ l HeLa nuclear extract (Computer Cell Culture Centre, Belgium) in 250 μ l IPH buffer²⁰ and incubated at 4 °C for 45 min. Beads were washed 2–3 times with IPH buffer and assayed for histone deacetylase activity. For peptide and E7 competition experiments, GST fusion proteins were preincubated with competitors for 5 min at room temperature before adding to the extract. Immunoprecipitations were carried out using standard procedures¹⁶. Antibody complexes were collected on protein A/G Sepharose beads, washed 3–4 times with IPH buffer and assayed for histone deacetylase activity.

Immunoprecipitations and western blot analysis. Standard procedures were used for coimmunoprecipitation and western blotting¹⁶. For coimmuno-

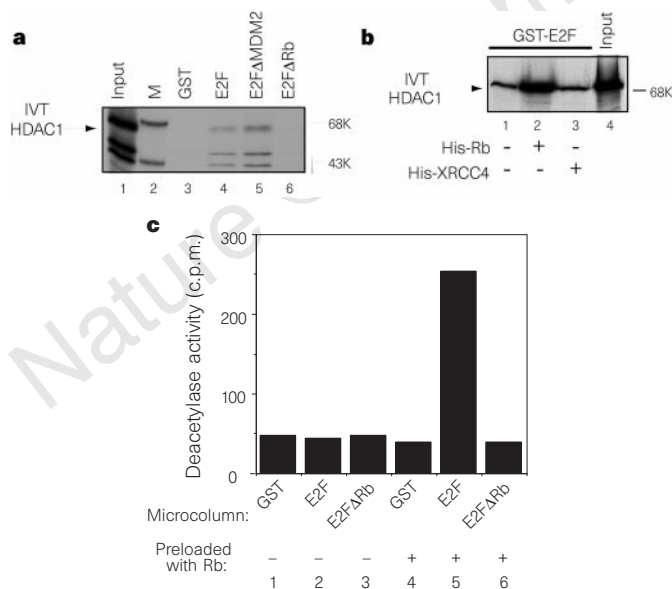


Figure 3 Rb recruits HDAC1 to E2F. **a**, Interaction of E2F and HDAC1 is dependent on the Rb-binding site of E2F. Binding of GST-E2F380–437 and GST-E2F mutants to *in vitro*-translated HDAC1 was assessed in a GST pull-down assay as indicated. Lane 1: ³⁵S-HDAC1 input (10%). E2F: GST-E2F380–437. E2FΔM2 contains a mutated Mdm2-binding site, E2FΔRb a mutated Rb-binding site. **b**, Binding of E2F to HDAC1 *in vitro* is enhanced by addition of recombinant Rb protein. Binding of *in vitro*-translated HDAC1 to GST-E2F380–437 was tested in the presence and absence of recombinant purified histidine-tagged Rb379–928 and XRCC4 protein, respectively, as indicated. Lane 4: ³⁵S-HDAC1 input (10%). **c**, Affinity purification of histone deacetylase activity from nuclear extract by a GST-E2F column requires Rb. GST, GST-E2F380–437 and GST-E2FΔRb microcolumns were preloaded with recombinant His-tagged Rb379–928 protein and used to bind deacetylase activity from nuclear extract.

precipitation of Rb and HDAC1, whole-cell extract from 10⁹ Jurkat cells in 10 ml IPH buffer was used.

Cell culture, transfection, CAT and luciferase assays. Cell lines were maintained in DMEM supplemented with 10% fetal calf serum and grown at 37 °C, 5% CO₂. The 3T3/208 cell line was maintained in the presence of 0.5 mg ml⁻¹ G418. Jurkat cells were maintained in RPMI medium supplemented

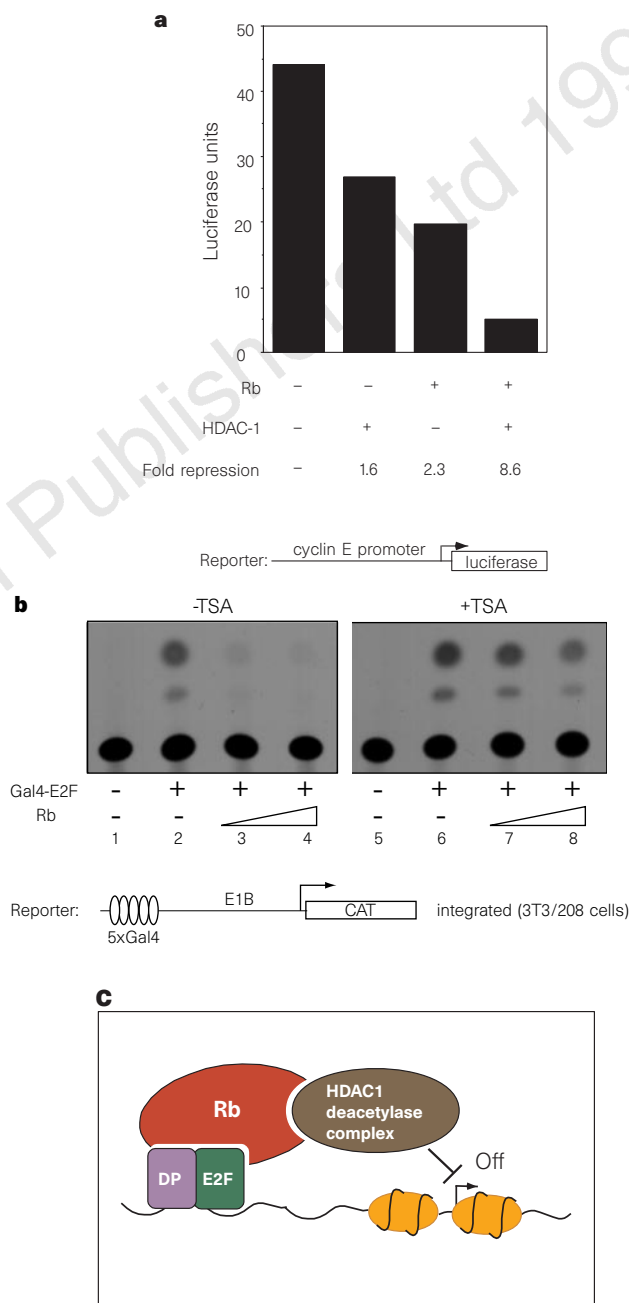


Figure 4 Histone deacetylase is required for full Rb-mediated transcriptional repression *in vivo*. **a**, HDAC1 enhances Rb-mediated repression of E2F *in vivo*. U2OS cells were transfected with pCE(-543/+263)luc reporter (1 μ g), pCMV-Rb379–928 (0.25 μ g) and pcDNA3-HDAC1-F (0.1 μ g) as indicated. **b**, Inhibition of histone deacetylase activity compromises repression of E2F by Rb *in vivo*. 3T3/208 cells containing a chromosomal Gal4-dependent CAT reporter gene (5GE1B-CAT) (ref. 19) were transfected with pHKG-E2F380–437 (10 μ g) and increasing amounts of pCMV-Rb379–928 (lanes 3 and 7: 0.3 μ g, lanes 4 and 8: 1.0 μ g) expression vectors in the absence (left) or presence (right) of 330 nM TSA as indicated. **c**, Repression of E2F by Rb involves deacetylase activity. Rb can bind simultaneously to E2F and HDAC1 and uses deacetylase activity to silence the promoter. This mechanism may account for the 'active' repression of Gal4-Rb and the E2F-Rb complex.

with 10% fetal calf serum. Transfections, CAT and luciferase assays were done using standard techniques¹⁵. TSA (Wako BioProducts) was added at 330 nM 24 h before collecting cells.

Recombinant GST fusion and histidine-tagged proteins. Recombinant proteins were expressed in and purified from *E. coli* XA90 as described²⁰. For competition experiments, GST-E7 proteins were cleaved from GST using thrombin.

GST microcolumns, pull-down assays and *in vitro* translation. Microcolumns were prepared by placing a glass bead in a yellow Gilson tip and applying 30–50 μ l GST fusion protein prebound to glutathione beads. GST-E2F380–437 columns were loaded with recombinant Rb protein and washed 4 times with IPH buffer. Beads were then transferred to a fresh reaction tube for affinity purification of histone deacetylase activity. *In vitro* translation reactions and GST pull-down experiments have been described¹⁵. HDAC1 was translated *in vitro* from pING14A-HDAC1. The sequences of the competitor peptides were LQPGTTDLYCYEQLNDSS (LXCXE) and YPYDVDPYA (HA), respectively.

Plasmids. pING14A-HDAC1 was subcloned from pBJ5-HD1-F by PCR (ref. 4). pGEX-Rb, pGEX-E7 and pGEX-E2F expression vectors, pCMV-Rb379–928, pHKG-E2F380–437 and pCycE-luc have all been described^{15–17}.

Received 12 August; accepted 19 December 1997.

- Weinberg, R. A. The retinoblastoma protein and cell cycle control. *Cell* **81**, 323–330 (1995).
- Weintraub, S. J., Prater, C. A. & Dean, D. C. Retinoblastoma protein switches the E2F site from positive to negative elements. *Nature* **358**, 259–261 (1992).
- Weintraub, S. J. *et al.* Mechanism of active transcriptional repression by the retinoblastoma protein. *Nature* **375**, 812–815 (1995).
- Taunton, J., Hassig, C. A. & Schreiber, S. L. A mammalian histone deacetylase related to the yeast transcriptional regulator Rpd3p. *Science* **272**, 408–411 (1996).
- Yang, W.-M., Inouye, C., Zeng, Y., Bearss, D. & Seto, E. Transcriptional repression by YY1 is mediated by interaction with a mammalian homolog of the yeast global regulator RPD2. *Proc. Natl Acad. Sci. USA* **93**, 12845–12850 (1996).
- Heinzel, T. *et al.* A complex containing N-CoR, mSin3 and histone deacetylase mediates transcriptional repression. *Nature* **387**, 43–48 (1997).
- Alland, L. *et al.* Role for N-CoR and histone deacetylase in Sin3-mediated transcriptional repression. *Nature* **387**, 49–55 (1997).
- Hassig, C. A., Fleischer, T. C., Billin, A. N., Schreiber, S. L. & Ayer, D. E. Histone deacetylase activity is required for full transcriptional repression by mSin3A. *Cell* **89**, 341–347 (1997).
- Laherty, C. D. *et al.* Histone deacetylases associated with the mSin3 corepressor mediate Mad transcriptional repression. *Cell* **89**, 349–356 (1997).
- Zhang, Y., Iatni, R., Erdjument-Bromage, H., Tempst, P. & Reinberg, D. Histone deacetylases and SAP18, a novel polypeptide, are components of a human Sin3 complex. *Cell* **89**, 357–364 (1997).
- Nagy, L. *et al.* Nuclear receptor repression mediated by a complex containing SMRT, mSin3A, and histone deacetylase. *Cell* **89**, 373–380 (1997).
- Kadosh, D. & Struhl, K. Repression by Ume6 involves recruitment of a complex containing Sin3 corepressor and Rpd3 histone deacetylase to target promoters. *Cell* **89**, 365–371 (1997).
- Qian, Y.-W. *et al.* A retinoblastoma-binding protein related to a negative regulator of Ras in yeast. *Nature* **364**, 648–652 (1993).
- Huang, S., Lee, W.-H. & Lee, E. Y.-H. P. A cellular protein that competes with SV40 T antigen for binding to the retinoblastoma gene product. *Nature* **350**, 160–162 (1991).
- Hagemeier, C., Cook, A. & Kouzarides, T. The retinoblastoma protein binds E2F residues required for activation *in vivo* and TBP binding *in vitro*. *Nucleic Acids Res.* **21**, 4998–5004 (1993).
- Martin, K. *et al.* Stimulation of E2F1/DP1 transcriptional activity by Mdm2 oncoprotein. *Nature* **375**, 691–694 (1995).
- Botz, J. *et al.* Cell cycle regulation of the murine cyclin E gene depends on an E2F binding site in the promoter. *Mol. Cell Biol.* **16**, 3401–3409 (1996).
- Hurford, R. K. Jr, Cobrinik, D., Lee, M. H. & Dyson, N. pRB and p107/p130 are required for the regulated expression of different sets of E2F responsive genes. *Genes Dev.* **11**, 1447–1463 (1997).
- Alevizopoulos, A. *et al.* A proline-rich TGF- β -responsive transcriptional activator interacts with histone H3. *Genes Dev.* **9**, 3051–3066 (1995).
- Bannister, A. J. & Kouzarides, T. The CBP co-activator is a histone acetyltransferase. *Nature* **384**, 641–643 (1996).
- Ogryzko, V. V., Schiltz, R. L., Russanova, V., Howard, B. H. & Nakatani, Y. The transcriptional coactivators p300 CBP are histone acetyltransferases. *Cell* **87**, 953–959 (1996).
- Gu, W. & Roeder, R. G. Activation of p53 sequence-specific DNA binding by acetylation of the p53 C-terminal domain. *Cell* **90**, 595–606 (1997).
- Cavanaugh, A. H. *et al.* Activity of RNA polymerase I transcription factor UBF blocked by Rb gene product. *Nature* **374**, 177–180 (1995).
- White, R. J., Trouche, D., Martin, K., Jackson, S. P. & Kouzarides, T. Repression of RNA polymerase III transcription by the retinoblastoma protein. *Nature* **382**, 88–90 (1996).
- Larminie, C. G. *et al.* Mechanistic analysis of RNA polymerase III regulation by the retinoblastoma protein. *EMBO J.* **16**, 2061–2071 (1997).

Acknowledgements. This work is dedicated to the memory of Danny Spiegelhalter who died of retinoblastoma. We thank C. Hassig and S. L. Schreiber for the 1121 and 1123 HCAC1 antibodies and for technical advice on peptide acetylation and histone deacetylase assay; S. Critchlow for the gift of XRCC4 protein; W. Zwierschke and P. Janssen-Duerr for the CycE reporter construct; N. Mermod for the 3T3/208 cell line and M. O'Connor for technical advice on GST microcolumns. This work was funded by a Cancer Research Campaign programme grant, by an MRC grant and by an EC Biomed2 grant. A.B. was supported by a grant from the Association for International Cancer Research; D.J.McC. was funded by an American Cancer Society International Cancer research fellowship. E.A.M. was supported by a CRC studentship; J.L.R. was funded by an MRC studentship.

Correspondence and requests for materials should be addressed to T.K. (e-mail: tk106@mole.bio.cam.ac.uk).

Retinoblastoma protein represses transcription by recruiting a histone deacetylase

L. Magnaghi-Jaulin*, R. Groisman*, I. Naguibneva*, P. Robin*, S. Lorain†, J. P. Le Villain‡, F. Troalen‡, D. Trouche* & A. Harel-Bellan*

* Laboratoire Oncogénèse, Différenciation et Transduction du Signal, CNRS UPR 9079, IFC-01, 94801 Villejuif, France

† CNRS URA 1156, ‡ Unité des Marqueurs Biologiques et Moléculaires, Institut Gustave Roussy, 94805 Villejuif, France

The retinoblastoma tumour-suppressor protein Rb¹ inhibits cell proliferation by repressing a subset of genes that are controlled by the E2F family of transcription factors² and which are involved in progression from the G1 to the S phase of the cell cycle. Rb, which is recruited to target promoters by E2F1 (ref. 3), represses

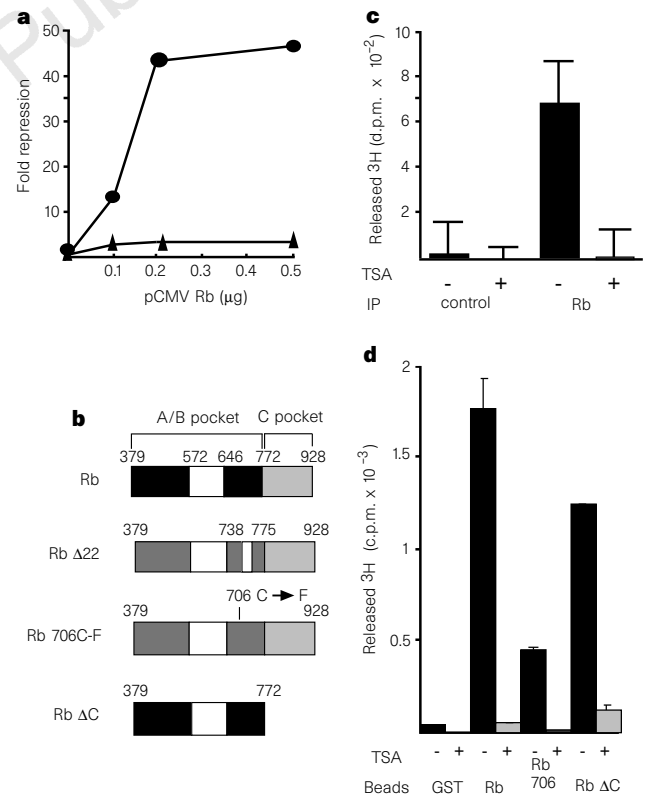


Figure 1 A histone deacetylase activity is associated with Rb. **a**, Trichostatin A (TSA) treatment relieves Rb repression of E2F1. SAOS2 cells were transiently transfected with 1 μ g pE2F-luc reporter vector, 50 ng pCMV E2F1 (ref. 19), and the indicated doses of pCMV Rb. Cells were treated (triangles) or not (circles) with 100 ng ml⁻¹ TSA for 8 h; repression by pCMV Rb was calculated. **b**, Diagram of the Rb mutants used. The Rb A/B pocket is indicated by filled boxes, and the C pocket by grey boxes. pRb Δ 22 lacks exon 22 and Rb706C-F contains a point mutation at codon 706; both mutations result in inactivation of the A/B pocket. pRb Δ C lacks the C pocket. **c**, A histone deacetylase is immunoprecipitated (IP) by anti-Rb antibodies: Jurkat whole-cell extracts were immunoprecipitated with anti-Rb or control (anti-HA) antibodies and immunoprecipitates were assayed for deacetylase activity. **d**, Recombinant Rb binds to a histone deacetylase activity: beads covered with the indicated GST fusion proteins were incubated with nuclear extracts (1 mg) and assayed for histone deacetylase activity.

transcription by masking the E2F1 transactivation domain⁴ and by inhibiting surrounding enhancer elements⁵⁻⁸, an active repression that could be crucial for the proper control of progression through the cell cycle⁹. Some transcriptional regulators act by acetylating or deacetylating the tails protruding from the core histones¹⁰, thereby modulating the local structure of chromatin: for example, some transcriptional repressors function through the recruitment of histone deacetylases¹¹. We show here that the histone deacetylase HDAC1 physically interacts and cooperates with Rb. In HDAC1, the sequence involved is an LXCXE motif, similar to that used by viral transforming proteins to contact Rb. Our results strongly suggest that the Rb/HDAC1 complex is a key element in the control of cell proliferation and differentiation and that it is a likely target for transforming viruses.

An inhibitor of histone deacetylases, trichostatin A (TSA)¹², was used to monitor the involvement of these enzymes in transcriptional repression by Rb by using transient transfection experiments

in Rb-negative cells. In the absence of TSA, optimal doses of Rb resulted in a 40–50-fold repression of a co-transfected luciferase reporter gene driven by E2F elements and transactivated by ectopically expressed E2F1. Treatment with TSA drastically reduced the inhibiting effect of Rb (Fig. 1a). TSA-insensitive repression (2–3 fold) is probably due to masking of the E2F1 transactivation domain. Rb includes two essential domains that are both required for growth inhibition, the A/B pocket and the C pocket (Fig. 1b). The A/B pocket is frequently mutated in tumours and is the site of interaction with transforming viral proteins; it is both necessary and sufficient for transcriptional repression^{6,8,13,14}. RbΔ22, a naturally occurring mutant of the A/B pocket (Fig. 1b), did not significantly repress E2F1-dependent activity, and TSA had no effect on this feeble repression (data not shown). TSA did not have any effect on the same reporter in the absence of E2F1, nor on a cytomegalovirus (CMV) promoter used as an internal control (data not shown); neither did it influence the expression or phosphorylation of

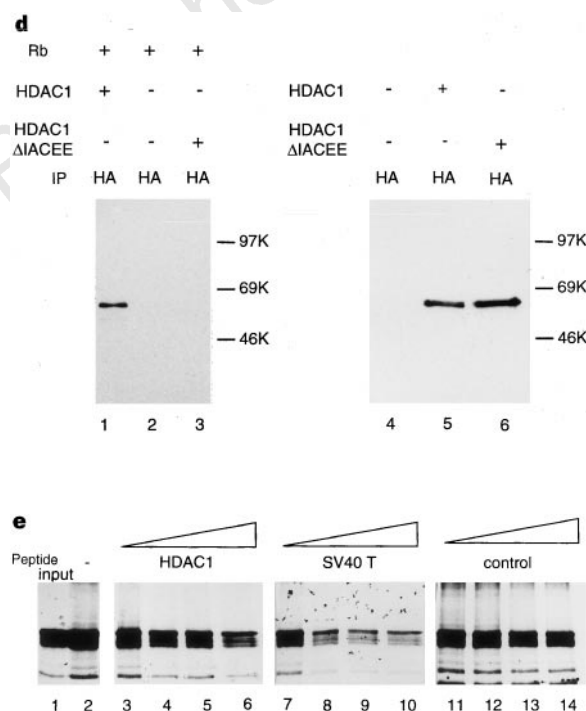
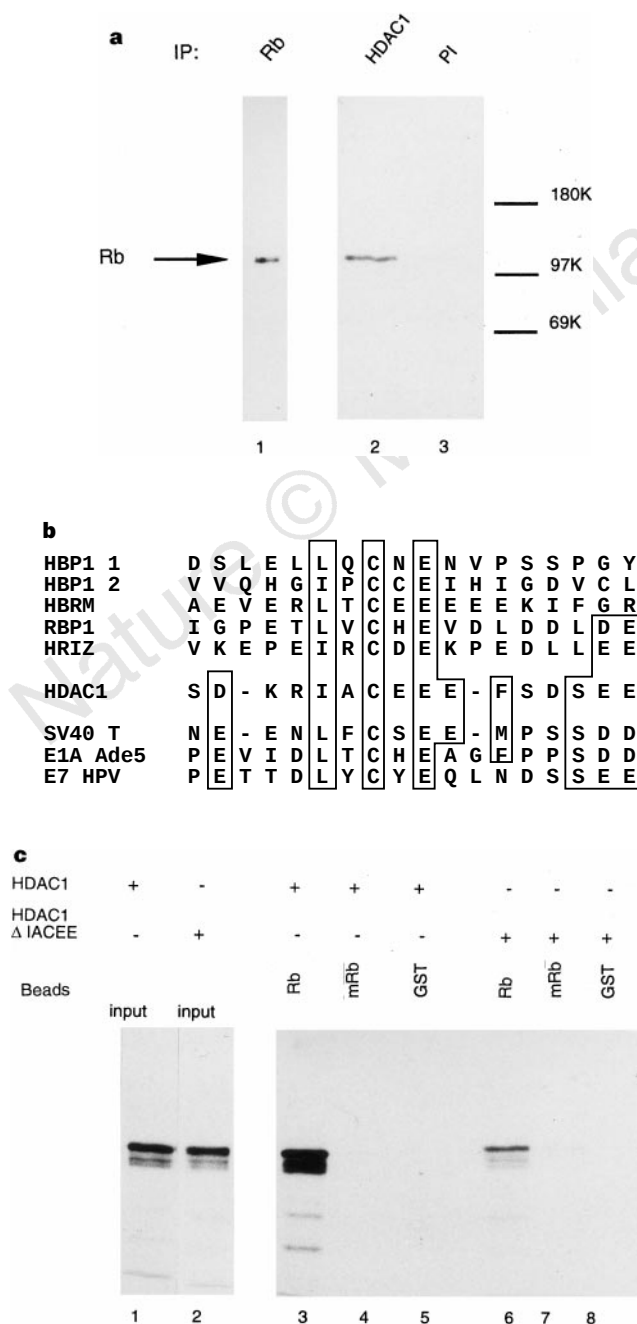


Figure 2 Rb interacts with HDAC1 through an 'LXCXE'-like motif. **a**, Rb and HDAC1 are part of a multimolecular complex: whole-cell extracts were immunoprecipitated with anti-Rb (lane 1), anti-HDAC1 (lane 2) or control (preimmune (PI) serum) antibodies (lane 3) and analysed by western blotting using anti-Rb antibodies. In lane 1, 1/30 of the immunoprecipitate was loaded. **b**, Alignment of HDAC1 amino acids from 410 to 425 with cellular (top) and viral (bottom) proteins interacting with Rb. Homologous amino acids are boxed. **c**, Involvement of the 'LXCXE'-like motif in Rb/HDAC1 interaction *in vitro*: labelled HDAC1 or HDAC1 ΔIACEE were subjected to a GST pull-down assay using beads covered with GST fusion proteins (mRb: Rb706 C-F). **d**, Involvement of the 'LXCXE'-like motif in Rb/HDAC1 interaction in live cells: extracts of cells transfected with 10 μg pCMV Rb and 10 μg of either pCMV HA-HDAC1 (lanes 1, 5), pCMV HA-HDAC1 ΔIACEE (lanes 3, 6), or pCMV HA empty vector (lanes 2, 4) were immunoprecipitated using anti-HA antibodies and analysed by western blot using anti-Rb (left) or anti-HA (right) antibodies. **e**, Inhibition of Rb/HDAC1 interaction by an excess of the 'LXCXE'-like motif: labelled HDAC1 was subjected to GST pull-down analysis using GST-Rb beads in the absence (lane 2) or in the presence of increasing doses (5, 10, 15, 20 μg) of a synthetic peptide corresponding to the HDAC1 motif (SDKRIA-CEEFSDSEE) (lanes 3–6), to the SV40T-antigen motif (NEENLFCSEEMPSSDD) (lanes 7–10) or to a control sequence (YVEPTKEYDVALKLDC) (lanes 11–14).

transfected Rb (data not shown). Taken together, our results indicate that a histone deacetylase is involved in the repression process.

One possibility is that the histone deacetylase involved in repression interacts with and is recruited by Rb. Immunoprecipitation of endogenous Rb from cell extracts results in the specific co-retention of a significant amount of TSA-sensitive histone deacetylase activity (Fig. 1c). In addition, beads covered with glutathione-S-transferase-Rb (GST-Rb) fusion proteins specifically retained large amounts of histone deacetylase activity (Fig. 1d) from similar extracts, indicating that an enzyme with this activity was specifically bound by the Rb fusion protein. Analysis of Rb mutants (Fig. 1b) indicated that binding essentially occurs through the repressor domain, the A/B pocket, although the C pocket could also be involved (Fig. 1d): binding decreased when the A/B pocket of Rb was mutated (Rb 706C-F), whereas deletion of the whole C pocket, which is dispensable for transcription repression, had little effect (Rb Δ C). These results show that the Rb repressor domain interacts with a histone deacetylase, either directly or through other molecules.

HDAC1 and HDAC2 are two mammalian proteins that show strong homology to the yeast histone deacetylase RPD3^{15,16}. HDAC1 interacts strongly with the Rb-binding protein RbA-p48 (ref. 16), which may physically link Rb with the histone deacetylases. When immunoprecipitated from cell extracts, HDAC1 precipitated with a significant amount of Rb protein (Fig. 2a, lane 2). This retention was specific as it was not observed with an irrelevant antibody (lane 3). Furthermore, in reciprocal experiments, HDAC1 was retained in Rb immunoprecipitates (data not shown). Taken together, these results suggest that Rb and HDAC1 exist in the cell as members of the same multimolecular complex. However, the interaction between HDAC1 and Rb was observed in GST pull-down assays using recombinant proteins (data not shown, and Fig. 2c), strongly suggesting that interaction is direct. The repressor domain of Rb seems to be responsible for this interaction, because HDAC1 was hardly recognized by a protein that carried a point mutation in the A/B pocket (706C-F; Fig. 1b) and could no longer repress transcription, whereas deletion of the C pocket had no effect on binding (data not shown).

HDAC1 contains an IACEE sequence (Fig. 2b) which is similar to the LXCXE or IXCXE motif through which several viral or cellular proteins interact with the A/B pocket of Rb (ref. 17) and which shows strong homology to the SV40 T-antigen sequence (Fig. 2b).

The HDAC1 IACEE motif is involved in the interaction with Rb, as binding of Rb to a mutant in which this sequence had been deleted was strongly decreased in GST pull-down assays (Fig. 2c: compare lanes 3 and 6). In addition, the same mutant was unable to retain Rb efficiently after co-immunoprecipitation from transfected cells (Fig. 2d). A synthetic peptide corresponding to the IACEE sequence was shown to inhibit specifically the interaction when tested by GST pull-down (Fig. 2e). A peptide harbouring the LXCXE motif of the SV40 T-antigen, one of the transforming viral proteins that release repression by Rb, was a better inhibitor than the homologous HDAC1 sequence. The interaction in the cell between HDAC1 and Rb might thus be targeted by viral transforming proteins such as SV40 T or E1A through their binding to Rb.

One possibility is that the HDAC1 protein is recruited to repressed promoters through interaction with Rb, and that the Rb molecule acts as a bridge between E2F1 and HDAC1. This requires Rb to bind to the transcription factor and HDAC1 simultaneously, which is not unreasonable given that Rb is able to bind E2F1 and several other LXCXE-containing proteins at the same time^{18,19}. We demonstrated this dual binding in a sandwich GST pull-down assay (Fig. 3): HDAC1 was specifically retained on E2F1-coated beads in the presence of recombinant Rb but not in its absence (Fig. 3: compare lanes 2 and 3); again, the HDAC1 Δ IACEE mutant molecule was not retained.

Our results from experiments using TSA suggest that a histone

deacetylase is involved in transcriptional repression by Rb, with at least one of the deacetylases being HDAC1. This was confirmed in transient transfection experiments (Fig. 4a): a limiting dose of Rb gave a moderate repression (3-fold as compared to 50-fold; with optimal doses of Rb) in the absence of HDAC1 the repression was increased when HDAC1 was co-expressed (Fig. 4a), which did not affect the expression of Rb (data not shown). At the doses used in these assays, HDAC1 expression did not repress E2F1 in the absence of Rb (Fig. 4a, b), nor did it repress an irrelevant CMV promoter

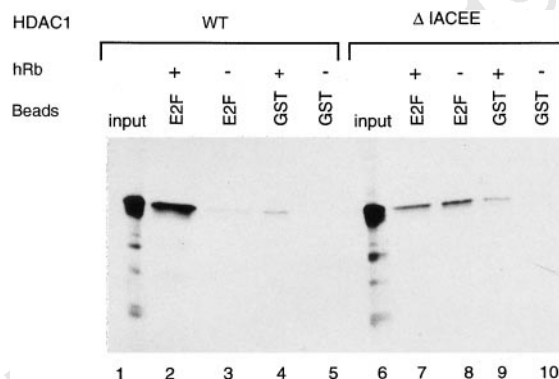


Figure 3 Rb interacts simultaneously with E2F1 and HDAC1. GST E2F1- or GST-coated beads were incubated with or without bacterially expressed His-Rb (hRb), as indicated. Beads were then used in a GST pull-down assay with labelled HDAC1 or HDAC1 Δ IACEE. WT, wild type.

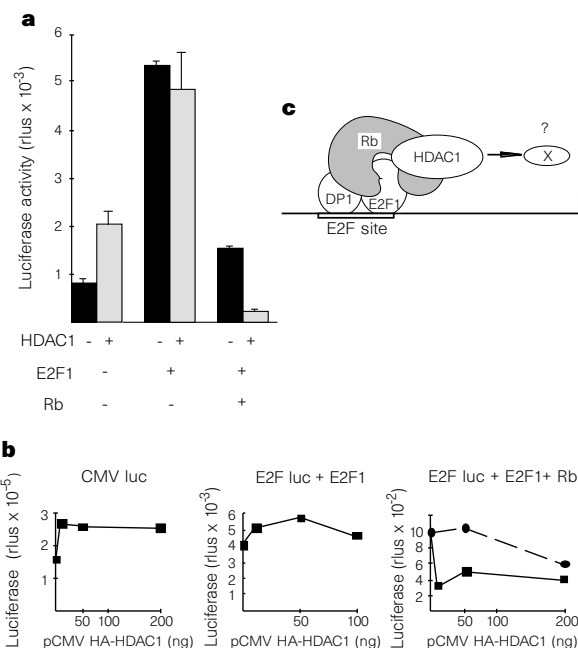


Figure 4 HDAC1 cooperates with Rb to repress E2F1 activity. **a**, SAOS2 cells were transiently transfected with 1 μ g E2F1-luciferase reporter vector and, where indicated, 50 ng pCMV E2F1 expression vector, 200 ng pCMV Rb, 2 μ g pBJ5 HDAC1-Flag¹⁶ (grey bars) or pBJ5 (solid bars) rlus: arbitrary light units. **b**, The IACEE motif is required for the cooperative effect. SAOS2 cells were transiently transfected with 1 μ g pCMV-luc (left), 1 μ g E2F-luc and 50 ng pCMV E2F1 (middle) or 1 μ g E2F-luc, 50 ng pCMV E2F1 and 200 ng pCMV Rb (right). Also cotransfected were the indicated amounts of pCMV HA-HDAC1 (squares with solid lines) or pCMV HA-HDAC1 Δ IACEE (circles, broken line). **c**, Repression by the E2F1/Rb complex. Rb recruits HDAC1 to promoters through E2F and allows HDAC1 to deacetylate a target protein (X) locally.

(Fig. 4b). Disruption of the interaction between Rb and HDAC1 by deletion of the IACEE motif decreased their cooperative effect on transcription repression (Fig. 4b).

The control of E2F1-regulated genes is thought to involve recruitment of Rb by E2F1 to the promoters of these genes, which first masks the transactivation domain of E2F1, thereby repressing E2F1, and second actively represses surrounding *cis*-acting elements or basal promoters through a poorly understood mechanism. Our results indicate that this active repression mechanism involves recruitment by E2F1 of a complex that includes Rb and HDAC1 (Fig. 4c); however, this recruitment only partly accounts for the repression mechanism, because TSA does not completely relieve the repression and because our preliminary results suggest that at least some endogenous E2F target genes are not induced by TSA (data not shown). The mechanism of repression is likely to be complex and masking of the E2F1 transactivation domain also plays a role. Histone acetylases or deacetylases are thought to modulate transcription by remodelling chromatin structure²⁰, but our results and those described for other repressors^{21–26} were obtained using transiently transfected plasmids, which do not seem to be arranged in a typical chromatin structure; therefore, HDAC1 may function by deacetylating proteins that are not histones, such as the HMG proteins or proteins of the basal transcriptional machinery.

Our results suggest that a complex between Rb and HDAC1 is recruited through E2F1 to promoters that are repressed during the G1 phase of the cell cycle. This complex is like the mSin3/HDAC1 complex recruited to promoters that are repressed by Mad/Mxi proteins^{21–24,26}, and like the NcoR/HDAC1 complex recruited by unoccupied nuclear receptors^{24,25}. Mad/Mxi, nuclear receptors and E2F all have two features in common: (1) they are switched, in response to external stimuli, from a repressive to an activating state; (2) they are involved in the control of cell proliferation and differentiation. HDAC1 might be specifically recruited to “dual” *cis*-acting elements on promoters that are subject to cell-cycle dependent regulation. □

Methods

Plasmids. Details of constructions are available upon request.

Transfections and luciferase assays. Transient transfections were carried out by calcium phosphate coprecipitation. Luciferase was assayed using reagents from Promega according to the manufacturer's instructions and β -galactosidase (from a CMV β -Gal, CAYLA, used as an internal control) was assayed using a kit from Tropic.

Immunoprecipitations. Anti-Rb antibodies were G3-245 or XZ55 (Pharmin-gen); anti-haemagglutinin (HA) antibody was 12AC5 (Boehringer) and anti-HDAC1 was 1123 antiserum (gift from S. Schreiber). For experiments on endogenous proteins, total cell extracts were prepared from 3×10^8 Jurkat cells. Cells were resuspended in 6 ml lysis buffer (50 mM Tris, pH 8.0, 300 mM NaCl, 10 mM MgCl₂, and protease inhibitors) supplemented with 0.4% N-P40 and incubated for 15 min on ice. The supernatant was cleared by centrifugation and mixed with 6 ml dilution buffer (50 mM Tris, pH 8.0) with 0.4% N-P40. Total cell extracts (30–60 mg) were precleared using 40 μ l protein A/protein G beads (Sigma), and immunoprecipitated by using standard procedures.

Transfected cells (COS) were resuspended in 500 μ l lysis buffer supplemented with 0.2% N-P40 and incubated for 15 min on ice. Supernatants were cleared by centrifugation, diluted using 500 μ l dilution buffer, and immunoprecipitated by using standard procedures.

Deacetylase assay. Deacetylase was assayed as described¹⁶, with incubation in TBS (20 mM Tris, pH 8.0, 150 mM NaCl) for 2 h at 37°C. All samples were assayed in duplicate.

GST pull-down assay. Beads coated with GST, GST–Rb, GST–Rb706C–F (ref. 27) or GST–E2F1 (ref. 28) were prepared as described²⁹. His–Rb was prepared using the Qiagen protocol. Beads were preincubated for 10 min at room temperature with rabbit reticulocyte lysate, then incubated for 1 h at room temperature with equivalent amounts of labelled *in vitro* translated HDAC1 or HDAC1– Δ IACEE proteins in a final volume of 20 μ l ELB buffer (250 mM NaCl, 50 mM HEPES, pH 7.9, 0.1% N-P40). For experiments on

ternary complexes, E2F1-coated beads or GST-coated control beads were incubated for 1 h at room temperature with soluble His–Rb or control in ELB buffer, washed once, then blocked with reticulocyte lysate and incubated for 1 h with equivalent amounts of *in vitro* translated ³⁵S-methionine-labelled HDAC1 or HDAC1– Δ IACEE in ELB buffer. Beads were washed five times in ELB buffer. Bound proteins were analysed on a 10% SDS–polyacrylamide gel and visualized either by using a PhosphorImager (Bas-1000, Fuji) or by conventional autoradiography.

Beads coated with GST, GST–Rb or GST–Rb706C–F were also used to retain a deacetylase activity from nuclear extracts prepared according to ref. 30. In this case, beads were blocked for 10 min with 20 μ l reticulocyte lysate, then incubated for 1 h at 4°C with nuclear extracts in buffer A (50 mM Tris, pH 8.0, 150 mM NaCl, 5 mM MgCl₂, 1 mM CaCl₂, 0.4% N-P40). Beads were then washed extensively with buffer A and assayed for deacetylase activity.

Western blot analysis. Western blots were made by using standard procedures with mouse monoclonal antibodies directed against E2F1 (KH75), Rb (XZ55 or G3-248) or the HA epitope (12CA5), and an ECL kit (Amersham).

Received 26 September; accepted 12 December 1997.

- Weinberg, R. A. The retinoblastoma gene and gene product. *Cancer Surv.* **12**, 43–57 (1992).
- Nevins, J. R. E2F: a link between the Rb tumor suppressor protein and viral oncoproteins. *Science* **258**, 424–429 (1992).
- Helin, K. *et al.* A cDNA encoding a pRB-binding protein with properties of the transcription factor E2F. *Cell* **70**, 337–350 (1992).
- Flemington, E. K., Speck, S. H. & Kaelin, W. G. Jr E2F-1-mediated transactivation is inhibited by complex formation with the retinoblastoma susceptibility gene product. *Proc. Natl Acad. Sci. USA* **90**, 6914–6918 (1993).
- Weintraub, S. J., Prater, C. A. & Dean, D. C. Retinoblastoma protein switches the E2F site from positive to negative element. *Nature* **358**, 259–261 (1992).
- Bremner, R. *et al.* Direct transcriptional repression by pRB and its reversal by specific cyclins. *Mol. Cell. Biol.* **15**, 3256–3265 (1995).
- Zacksenhaus, E., Jiang, Z., Phillips, R. A. & Gallie, B. L. Dual mechanisms of repression of E2F1 activity by the retinoblastoma gene product. *EMBO J.* **15**, 5917–5927 (1996).
- Chow, K. N., Starostik, P. & Dean, D. C. The Rb family contains a conserved cyclin-dependent-kinase-regulated transcriptional repressor motif. *Mol. Cell. Biol.* **16**, 7173–7181 (1996).
- Weinberg, R. A. E2F and cell proliferation: a world turned upside down. *Cell* **85**, 457–459 (1996).
- Turner, B. M. Histone acetylation and control of gene expression. *J. Cell Sci.* **99**, 13–20 (1991).
- Pazin, M. J. & Kadonaga, J. T. What's up and down with histone deacetylation and transcription? *Cell* **89**, 325–328 (1997).
- Yoshida, M., Kijima, M., Akita, M. & Beppu, T. Potent and specific inhibition of mammalian histone deacetylase both *in vivo* and *in vitro* by trichostatin A. *J. Biol. Chem.* **265**, 17174–17179 (1990).
- Weintraub, S. J. *et al.* Mechanism of active transcriptional repression by the retinoblastoma protein. *Nature* **375**, 812–815 (1995).
- Chow, K. N. & Dean, D. C. Domains A and B in the Rb pocket interact to form a transcriptional repressor motif. *Mol. Cell. Biol.* **16**, 4862–4868 (1996).
- Yang, W. M., Inouye, C., Zeng, Y., Bearss, D. & Seto, E. Transcriptional repression by YY1 is mediated by interaction with a mammalian homolog of the yeast global regulator RPD3. *Proc. Natl Acad. Sci. USA* **93**, 12845–12850 (1996).
- Taunton, J., Hassig, C. A. & Schreiber, S. L. A mammalian histone deacetylase related to the yeast transcriptional regulator Rpd3p. *Science* **272**, 408–411 (1996).
- Taya, Y. RB kinases and RB-binding proteins: new points of view. *Trends Biochem. Sci.* **22**, 14–17 (1997).
- Fattaey, A. R., Harlow, E. & Helin, K. Independent regions of adenovirus E1A are required for binding to and dissociation of E2F-protein complexes. *Mol. Cell. Biol.* **13**, 7267–7277 (1993).
- Trouche, D. *et al.* Hbrm is a downstream target for E2F1 repression by RB. *Proc. Natl Acad. Sci. USA* **94**, 11295–11300 (1997).
- Wolffe, A. P. & Pruss, D. Targeting chromatin disruption: Transcription regulators that acetylate histones. *Cell* **84**, 817–819 (1996).
- Alland, L. *et al.* Role for N-CoR and histone deacetylase in Sin3-mediated transcriptional repression. *Nature* **387**, 49–55 (1997).
- Zhang, Y., Iratni, R., Erdjument-Bromage, H., Tempst, P. & Reinberg, D. Histone deacetylases and SAP18, a novel polypeptide, are components of a human Sin3 complex. *Cell* **89**, 357–364 (1997).
- Laherty, C. D. *et al.* Histone deacetylases associated with the mSin3 corepressor mediate made transcriptional repression. *Cell* **89**, 349–356 (1997).
- Heinzel, T. *et al.* A complex containing N-CoR, mSin3 and histone deacetylase mediates transcriptional repression. *Nature* **387**, 43–48 (1997).
- Nagy, L. *et al.* Nuclear receptor repression mediated by a complex containing SMRT, mSin3A, and histone deacetylase. *Cell* **89**, 373–380 (1997).
- Hassig, C. A., Fleischer, T. C., Billin, A. N., Schreiber, S. L. & Ayer, D. E. Histone deacetylase activity is required for full transcriptional repression by mSin3A. *Cell* **89**, 341–347 (1997).
- Kaelin, W. G., Pallas, D. C., DeCaprio, J. A., Kaye, F. & Livingston, D. M. Identification of cellular proteins that can interact specifically with the T/E1A-binding region of the retinoblastoma gene product. *Cell* **64**, 521–532 (1991).
- Bandara, L. R. *et al.* DP-1: a cell cycle-regulated and phosphorylated component of transcription factor DRTF1/E2F which is functionally important for recognition by pRB and the adenovirus E4 orf 6/7 protein. *EMBO J.* **13**, 3104–3114 (1994).
- Groisman, R. *et al.* Physical interaction between the mitogen-responsive serum response factor and myogenic bHLH proteins. *J. Biol. Chem.* **271**, 5258–5264 (1996).
- Dignam, J. D., Lebowitz, R. M. & Roeder, R. G. Accurate transcription initiation by RNA polymerase II in a soluble. *Nucleic Acids Res.* **11**, 1474–1486 (1983).

Acknowledgements. This paper is dedicated to the memory of Jacques G. Harel who originated this work. We thank N. La Thangue, W. Kaelin, T. Kouzarides, C. Hassig, E. Lam, S. A. Leibovitch, R. Watson, S. Schreiber and S. Khochbin for materials, and C. Jaulin, G. Almouzni, S. Kochbin and L. Pritchard for

discussion. This work was supported by grants from the Ligue Nationale contre le Cancer, the Comité des Yvelines and the Comité du Val de Marne, from the Association pour la Recherche sur le Cancer and from the Groupement des Entreprises Françaises dans la Lutte contre le Cancer. L.M.J. held a fellowship from the Société Française du Cancer; R.G. held a fellowship from the Association Française contre les Myopathies.

Correspondence and requests for materials should be addressed to A.H.-B. (e-mail: ahbellan@vjf.cnrs.fr).

The conduction pore of a cardiac potassium channel

Kwok-Keung Tai & Steve A. N. Goldstein

Section of Developmental Biology and Biophysics, Departments of Pediatrics and Cellular and Molecular Physiology, Boyer Center for Molecular Medicine, Yale University School of Medicine, 295 Congress Avenue, New Haven, Connecticut 06536, USA

Ion channels form transmembrane water-filled pores that allow ions to cross membranes in a rapid and selective fashion. The amino acid residues that line these pores have been sought to reveal the mechanisms of ion conduction and selectivity^{1–7}. The pore (P) loop⁸ is a stretch of residues that influences single-channel-current amplitude, selectivity among ions and open-channel blockade^{2,3,5} and is conserved in potassium-channel sub-

units previously recognized to contribute to pore formation^{5,9}. To date, potassium-channel pores have been shown to form by symmetrical alignment of four P loops around a central conduction pathway^{10–12}. Here we show that the selectivity-determining pore region of the voltage-gated potassium channel of human heart through which the I_{Ks} current passes includes the transmembrane segment of the non-P-loop protein minK. Two adjacent residues in this segment of minK are exposed in the pore on either side of a short barrier that restricts the movement of sodium, cadmium and zinc ions across the membrane. Thus, potassium-selective pores are not restricted to P loops or a strict P-loop geometry.

In the human heart, two voltage-dependent potassium channels, carrying the currents I_{Ks} and I_{Kr} , open during muscle contraction to repolarize the myocardium and end each heartbeat¹³. I_{Ks} channels are formed by assembly of KvLQT1, a subunit with a single P loop, with minK, a 129-residue protein with a single transmembrane segment^{14,15}. HERG, the P-loop subunit that mediates I_{Kr} , also co-assembles with minK¹⁶. Controversy has surrounded the role of minK in these channels—some researchers argue that minK is a structural protein intrinsic to I_{Ks} -channel formation whereas others hold that it is an accessory subunit that functions as a channel regulator^{15–22}.

To determine whether minK sites are exposed in the I_{Ks} pore we performed scanning susceptibility analysis using the thiol-reactive transition-metal Cd^{2+} and 35 minK mutants, in each of which a

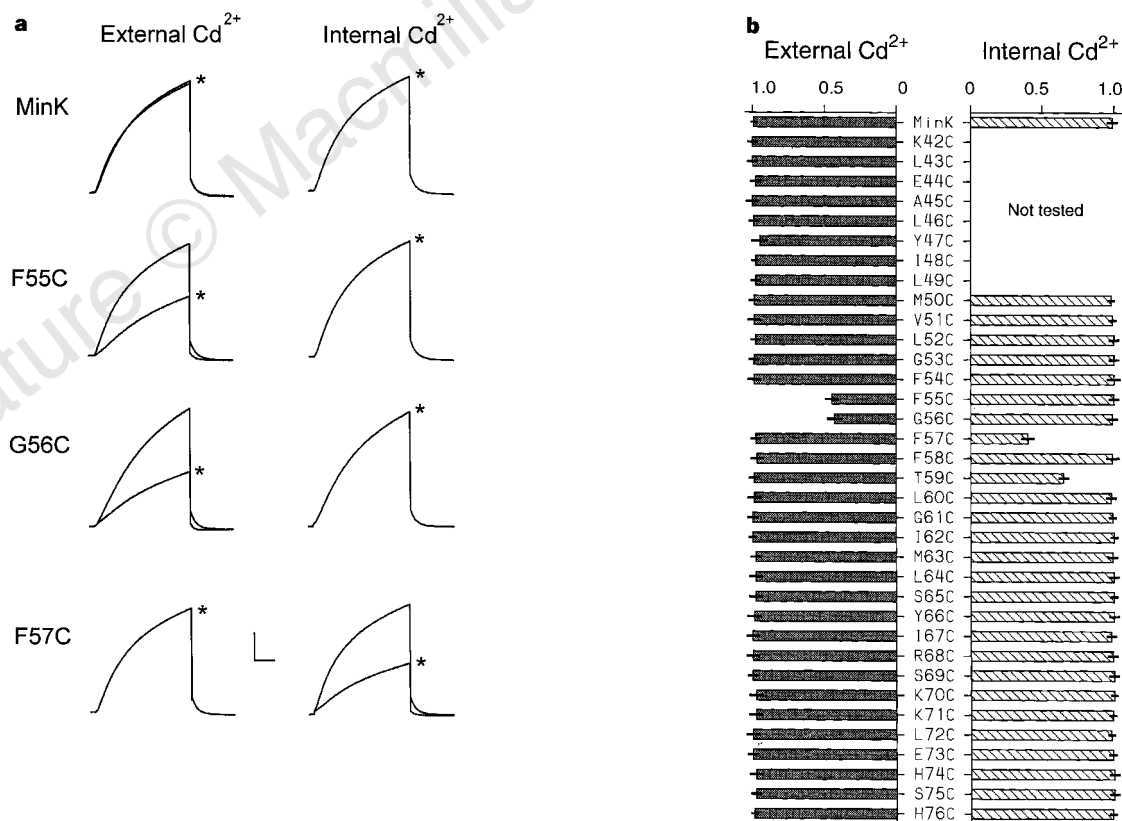


Figure 1 Blockade of channels formed with rat minK by Cd^{2+} . Whole-oocyte currents elicited by 10-s pulses from a holding potential of -80 mV to 0 mV with an interpulse interval of 20 s recorded by two-electrode voltage clamp are shown after leak subtraction. **a**, Current traces for sample oocytes before and after (asterisk) exposure to external Cd^{2+} (bath application of 0.5 mM $CdCl_2$ for 5 min) or internal Cd^{2+} (microinjection of 23 pmol $CdCl_2$; see Methods). Scale bars (bottom row) represent 200 nA (vertical bar) and 2 s (horizontal bar). **b**, The fraction of unblocked current at the end of a 10 s test pulse as described above for each

mutant studied (mean $\pm$ s.e.m. for 5 – 20 oocytes). cRNAs encoding minKs with cysteine residues at positions 62 , 63 , 65 , 66 , 69 , 70 and 71 were co-injected with KvLQT1 cRNA (see Methods). Calculated inhibition constants were 0.035 ± 0.003 mM for F55C, 0.11 ± 0.01 mM for G56C and roughly 0.1 and 0.2 mM for F57C and T59C, as estimated previously²⁰. Internal Cd^{2+} produces a voltage-independent open state and decreased time-dependent currents through T59C channels (not shown).

single residue was mutated to cysteine in sequential positions along the transmembrane segment^{6,20,23}. In the preliminary screen, we evaluated channels composed of rat minK and a subunit native to *Xenopus* oocytes^{14,19,21,24}. Wild-type minK (Fig. 1) and a mutant in which the single native cysteine in minK was altered to alanine (C107A, not shown) were both insensitive to Cd^{2+} , whether the reagent was applied extracellularly or microinjected into the cytosol.

In contrast, serial mutation of residues 42–76 to cysteine produced two channels that were sensitive to external Cd^{2+} but remained insensitive to internal reagent; these channels carried F55C and G56C mutations (Fig. 1a, b). Two other channels, with F57C or T59C mutations, were sensitive to internal but not external reagent (Fig. 1a, b). Apparently, Cd^{2+} applied from the external side of the membrane could not move past residue G56 to reach position

57. Conversely, Cd^{2+} applied internally could interact with residue 57 but was unable to traverse the pore to reach position 56.

This pattern of reactivity was also observed for I_{Ks} channels formed with human minK (h-minK) when equivalent residues were mutated to cysteine. Both I_{Ks} channels (formed with wild-type h-minK and KvLQT1) and channels formed with KvLQT1 alone were insensitive to Cd^{2+} , whether applied from the external or internal solution (Fig. 2a, b). However, channels formed with F54C or G55C mutant h-minK were sensitive to external but not internal Cd^{2+} , whereas those with mutation F56C were sensitive to internal but not external reagent (Fig. 2a, b). The theory that there is a barrier to Cd^{2+} movement between residues G55 and F56 in the human I_{Ks} -channel pore is tenable only if Cd^{2+} blocks such movement by binding in and occluding the I_{Ks} -conduction pathway.

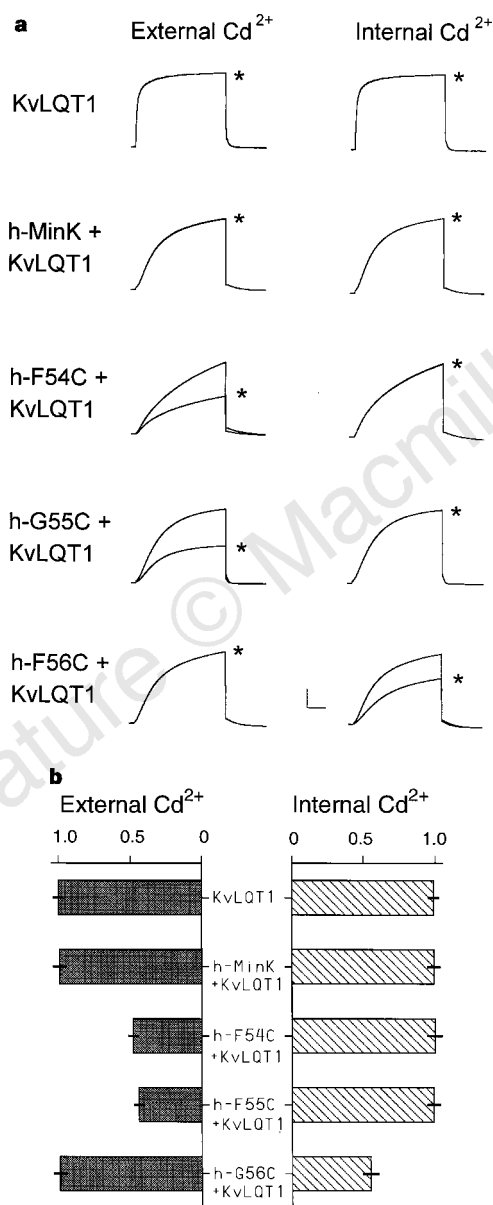


Figure 2 Blockade by Cd^{2+} of channels formed with KvLQT1 alone or with human minK and KvLQT1. **a**, Current traces for sample oocytes before and after (asterisk) exposure to external Cd^{2+} or internal Cd^{2+} , as in Fig. 1; scale bars represent 200 nA (vertical bar) and 2 s (horizontal bar). **b**, The fraction of unblocked current (mean $\pm$ s.e.m. for four to six oocytes) as in Fig. 1. Calculated inhibition constants were 0.11 ± 0.05 mM for h-F54C + KvLQT1, 0.32 ± 0.02 mM for h-G55C + KvLQT1 and ~ 0.3 mM for h-F56C + KvLQT1, as estimated previously²⁰.

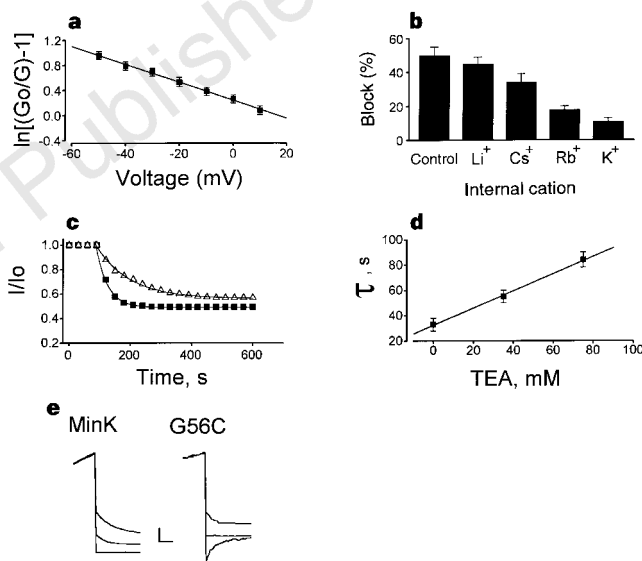


Figure 3 Inhibition of G56C rat minK channels by external Cd^{2+} shows the hallmarks of pore blockade. **a**, Blockage of G56C minK channels by external Cd^{2+} is voltage-sensitive; channels were opened by a 10-s command pulse to +20 mV and current was measured 25 ms after repolarization to test potentials between -50 and +10 mV; G_0 and G , macroscopic conductances in the presence and absence of 0.5 mM Cd^{2+} , respectively (mean $\pm$ s.e.m. for four oocytes). **b**, Block of G56C channels by external Cd^{2+} is altered by internal permeant cations; the fraction of blocked current by 0.5 mM external Cd^{2+} is shown before and after oocytes were injected with 20 nmol of the indicated salt solution; whole-cell currents were measured as in Fig. 1. **c**, External TEA slows the rate of block of G56C channels by external Cd^{2+} . The current was measured as in Fig. 1, first with constant flow of ND-96 (see Methods) and then during solution exchanges to either 0.5 mM Cd^{2+} (solid squares) or 75 mM TEA solution followed by 0.5 mM Cd^{2+} with 75 mM TEA (open triangles). The curves are single-exponential fits to the Cd^{2+} -inhibition time course. **d**, The effect of TEA on the time constant (τ) for Cd^{2+} -mediated inhibition. The time constant for block in the absence of TEA is 39 ± 3 s. **e**, G56C minK channels are permeable to Na^+ ions whereas wild-type channels are not. Tail current reversal potentials (V_{rev}) were determined by opening channels with a 10-s pulse to +20 mV and measuring current 50 ms after shifting to test potentials from -70 to -150 mV in 40-mV steps; scale bars represent 200 nA (vertical bar) and 25 ms (horizontal bar). Permeability ratios were calculated according to $P_{\text{K}}/P_{\text{Na}} = \exp(-FV_{\text{rev}}/RT)$ where R , T and F have their usual values and P_{K} and P_{Na} are measures of the permeability of K^+ and Na^+ ions, respectively. Wild-type V_{rev} is -150 ± 5 and 1 ± 1 in 98 mM NaCl and KCl solutions, respectively, whereas G56C V_{rev} is -107 ± 5 and 1 ± 1 , giving a permeability ratio $P_{\text{K}}/P_{\text{Na}}$ for wild-type channels of >350 and of ~ 50 for G56C channels. Values are mean $\pm$ s.e.m. for 5–7 oocytes.

Four characteristics of the inhibition by external Cd^{2+} indicate that it acts by a pore-blocking mechanism. First, blockade of G56C minK channels by external Cd^{2+} was voltage-dependent (Fig. 3a), as expected for a charged blocker that moves into the transmembrane electric field¹⁷. The block was rapidly reversible at this site and exhibited a δ value of 0.20 ± 0.03 ($n = 5$ oocytes), where δ is the fraction of the applied voltage drop experienced at the binding site. Second, block by external Cd^{2+} was sensitive to levels of permeant ions on the inside of the membrane (Fig. 3b). Raising intracellular levels of monovalent cations revealed that potassium or rubidium decreased the binding affinity of external Cd^{2+} at position G56C, whereas caesium had only a small effect and lithium was ineffective. Each cation decreased the affinity of external Cd^{2+} for the pore according to its position in the relative permeability series for open minK channels¹⁷, consistent with the notion that the internal test cations enter and traverse the conduction pore to destabilize Cd^{2+} on its external site^{20,25}.

Third, the kinetics of inhibition by Cd^{2+} in the absence and presence of external tetraethylammonium (TEA) also support a pore-blocking mechanism (Fig. 3c). External TEA is a weak but well characterized blocker of the I_{Ks} pore^{17,19,20}. As external TEA concentration was increased, the time constant (τ) of blockage by Cd^{2+} increased in a linear fashion directly proportional to the fraction of channels that were not blocked by TEA at equilibrium (Fig. 3d). This is expected for competitive inhibition by agents that reversibly block channels at overlapping binding sites²⁰. Finally, mutation of position 56 altered the ion-selectivity characteristics of the I_{Ks} channel. Mutant channels became permeable to Na^+ ions; this was revealed by inward sodium currents that were absent in wild-type channels (Fig. 3e). This change in selectivity was associated with a greater than five-fold increase in relative Na^+ permeability, as assessed by reversal potential measurements^{17,21}.

As the release of bound Cd^{2+} from the inside of F57C channels proceeds slowly (not shown), its use as a probe was limited. Zn^{2+} ,

however, blocked F57C channels from the intracellular but not external solution, did not block wild-type channels, and acted in a readily reversible fashion (Fig. 4a). As expected for an agent that acts by a pore-occlusion mechanism, blockade of F57C channels by internal Zn^{2+} exhibited both voltage dependence ($\delta = 0.33 \pm 0.06$ from the inside; $n = 6$ oocytes; Fig. 4b) and sensitivity to external monovalent cations according to the known permeability of the channel (Fig. 4c). Cysteine substitution of minK at positions 56 and 57 creates binding sites in the pore for external Cd^{2+} and internal Zn^{2+} , respectively, and neither metal can move inwards past G56C or outwards past F57C. This identifies these minK sites as boundaries of a selectivity barrier in the deep I_{Ks} pore.

The idea that a short region in the I_{Ks} pore selects against passage of Na^+ , Cd^{2+} and Zn^{2+} ions corresponds with the hypothesis that efficient movement of charges across cell membranes involves the limitation of unfavourable interactions of charges and low dielectric regions by close apposition of the bulk solutions on either side of the membrane across a short restrictive pathway^{26–28}. The conduction pathways of *Shaker*^{6,7} and Kv2.1 (ref. 29) potassium channels may also have short barriers, as judged by the residues that are exposed in these pathways. Although the physical barrier to permeation in minK appears to be short, as much as half the transmembrane electric field is traversed between minK residues 56 and 57. Thus, external Cd^{2+} moves ~20% into the field to bind at position 56 (Fig. 3) whereas internal Zn^{2+} moves outward through ~30% of the field to position 57 (Fig. 4). These data do not differentiate between a molecular-sieve or a binding-site mechanism for selective permeation^{26,30}. Although the atomic diameter of impermeant Cd^{2+} is ~1.41 Å and that of K^+ is ~2.66 Å, Cd^{2+} will be more heavily hydrated.

The transmembrane segment of minK has previously been detected in the external I_{Ks} pore²⁰. Here we show that this segment lines the deep pore. We are now working to establish the role of this segment in the formation of the cytoplasmic portion of the I_{Ks} conduction pathway and have succeeded in identifying minK sites that mediate blockade by internal TEA (K.K.T., F. Sesti and S.A.N.G., unpublished observations). These findings refute the theory that minK is simply a channel regulator and that its transmembrane segment is an inert membrane tether²². The data do not indicate the subunit stoichiometry of complete I_{Ks} channels—an intriguing issue given evidence for two minK monomers in complexes formed with the subunit endogenous to *Xenopus* oocytes¹⁹ and the likelihood that channels composed of KvLQT1 subunits in the absence of minK are tetrameric. The results also do not indicate whether P loops maintain a single conformation in different pores or modify their structure or function. That minK can co-assemble with P-loop subunits of at least two subfamilies (KvLQT1 and HERG¹⁶) argues such heteromultimeric pores will be found in other channels and reveals another mechanism for generation of potassium-channel diversity. □

Methods

Molecular biology. Mutants of rat and human minK were made and sequenced as previously described^{20,21}. All cysteine mutants of rat minK were otherwise cysteine-free (C107A). Human minK and human KvLQT1 cDNAs were gifts from R. Swanson (Merck), and M. Keating and M. Sanguinetti (Utah)¹⁴, respectively. Complementary RNAs were transcribed from rat minK genes in pSD^{20,21} and human minK and KvLQT1 in pBF2 (ref. 12). Two or more days after injection of 2 ng rat minK cRNA or 1 ng human minK cRNA and/or 5 ng human KvLQT1 cRNA, whole-cell currents were measured by two-electrode voltage clamp under constant perfusion at 22 °C¹⁹. Data were filtered at 100 Hz and sampled at 1 kHz, and leak correction was performed off-line¹⁷. **Electrophysiology.** Bath solution was either ND-96, as follows (in mM): 96 NaCl, 2 KCl, 1 MgCl₂, 0.3 CaCl₂ and 5 HEPES at pH 7.6, or solutions in which NaCl and/or KCl was replaced isotonicly by RbCl, CsCl, LiCl, CdCl₂, ZnCl₂ or TEACl. To study the effects of intracellular blockers, oocytes were microinjected with 23 nl unbuffered aqueous solutions containing 23 pmol

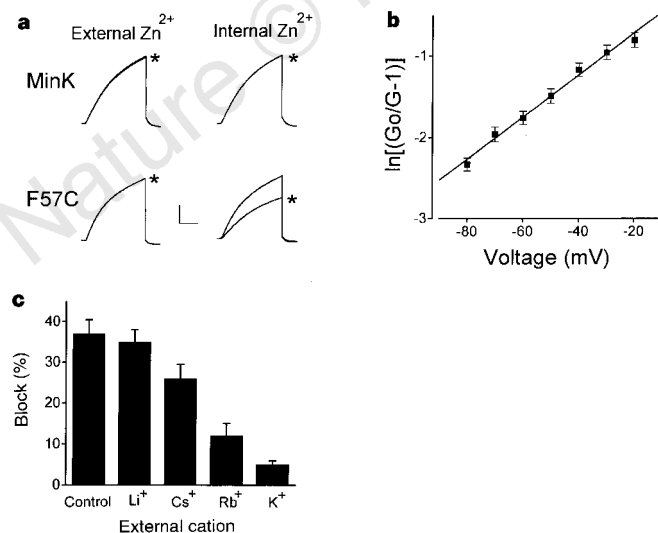


Figure 4 Inhibition of F57C rat minK channels by internal Zn^{2+} shows the hallmarks of pore blockade. **a**, Current traces for sample oocytes before and after (asterisk) exposure to external Zn^{2+} (bath application of 0.5 mM ZnCl_2 for 5 min) or internal Zn^{2+} (microinjection of 23 pmol ZnCl_2 ; see Methods) as in Fig. 1; scale bars represent 100 nA (vertical bar) and 3 s (horizontal bar). **b**, Block of F57C channels by internal Zn^{2+} is voltage-sensitive (protocol as in Fig. 3a; mean $\pm$ s.e.m. for four oocytes). **c**, Block of F57C channels by internal Zn^{2+} is altered by external permeant cations; bars show the fraction of blocked current when Zn^{2+} -injected oocytes were studied in bath solutions containing 100 mM of the indicated chloride salt; currents measured as in Fig. 1 (mean $\pm$ s.e.m. for four oocytes).

ZnCl₂ or CdCl₂, 2.3 nmol TEACl or 20 nmol of the chloride salt of monovalent cations, as described²⁰, and were studied 30–60 min thereafter.

Recorded currents showed no evidence that I_{Ks} channels contained the native oocyte subunit two days after minK and KvLQT1 cRNAs were co-injected: channels formed by injecting only minK cRNA were blocked when the native subunit was modified by methanethiosulfonate (MTS)-ethylammonium or MTS-ethyltrimethylammonium²¹, whereas those containing minK and KvLQT1 were MTS-insensitive (not shown). When minK and KvLQT1 were coexpressed, ~60% of the current could be blocked one day after cRNA injection; by day 2 and thereafter, all channels were insensitive to MTS reagents, indicating replacement of sensitive native subunits by KvLQT1.

Of 35 positions mutated to cysteine in rat minK, 7 (I62C, M63C, S65C, Y66C, S69C, K70C and K71C) did not express current when complexed with the native *Xenopus* oocyte subunit. These mutants could, however, co-assemble with KvLQT1 and gain current expression at the cell surface. Coexpression of each of the seven mutants with KvLQT1 altered the fast, saturating activation kinetics seen in channels formed by KvLQT1 alone to the slow, non-saturating kinetics characteristic of I_{Ks} channels¹⁴.

Received 22 September; accepted 17 November 1997.

- MacKinnon, R. & Miller, C. Mutant potassium channels with altered binding of charybdotoxin, a pore-blocking peptide inhibitor. *Science* **245**, 1382–1385 (1989).
- Yellen, G., Jurman, M. E., Abramson, T. & MacKinnon, R. Mutations affecting internal TEA blockage identify the probably pore-forming region of a K⁺ channel. *Science* **251**, 939–942 (1991).
- Hartmann, H. A. et al. Exchange of conduction pathways between two related K⁺ channels. *Science* **251**, 942–944 (1991).
- Goldstein, S. A., Pheasant, D. J. & Miller, C. The charybdotoxin receptor of a Shaker K⁺ channel: peptide and channel residues mediating molecular recognition. *Neuron* **12**, 1377–1388 (1994).
- Heginbotham, L., Lu, Z., Abramson, T. & MacKinnon, R. Mutations in the K⁺ channel signature sequence. *Biophys. J.* **66**, 1061–1067 (1994).
- Lu, Q. & Miller, C. Silver as a probe of pore-forming residues in a potassium channel. *Science* **268**, 304–307 (1995).
- Ranganathan, R., Lewis, J. H. & MacKinnon, R. Spatial localization of the K⁺ channel selectivity filter by mutant cycle-based structure analysis. *Neuron* **16**, 131–139 (1996).
- MacKinnon, R. Pore loops: an emerging theme in ion channel structure. *Neuron* **14**, 889–892 (1995).
- Goldstein, S. A. N., Wang, K. W., Ilan, N. & Pausch, M. Sequence and function of the two P domain potassium channels: implications of an emerging superfamily. *J. Mol. Med.* **76**, 13–20 (1998).
- MacKinnon, R. Determination of the subunit stoichiometry of a voltage-activated potassium channel. *Nature* **350**, 232–235 (1991).
- Shen, K. Z. et al. Tetraethylammonium block of Slowpoke calcium-activated potassium channels expressed in *Xenopus* oocytes: evidence for tetrameric channel formation. *Pflügers Arch.* **426**, 440–445 (1994).
- Glowatzki, E. et al. Subunit-dependent assembly of inward-rectifier K⁺ channels. *Proc. R. Soc. Lond. B* **261**, 251–261 (1995).
- Sanguinetti, M. C. & Jurkiewicz, N. K. Two components of cardiac delayed rectifier K⁺ current. Differential sensitivity to block by class III antiarrhythmic agents. *J. Gen. Physiol.* **96**, 195–215 (1990).
- Sanguinetti, M. C. et al. Coassembly of K(V)LQT1 and MinK (Isk) proteins to form cardiac I-Ks potassium channel. *Nature* **384**, 80–83 (1996).
- Barhanin, J. et al. K(V)LQT1 and Isk (minK) proteins associate to form the I(Ks) cardiac potassium current. *Nature* **384**, 78–80 (1996).
- McDonald, T. V. et al. A minK-HERG complex regulates the cardiac potassium current I_{Kr}. *Nature* **388**, 289–292 (1997).
- Goldstein, S. A. & Miller, C. Site-specific mutations in a minimal voltage-dependent K⁺ channel alter ion selectivity and open-channel block. *Neuron* **7**, 403–408 (1991).
- Attali, B. et al. The protein Isk is a dual activator of K⁺ and Cl⁻ channels. *Nature* **365**, 850–852 (1993).
- Wang, K. W. & Goldstein, S. A. N. Subunit composition of minK potassium channels. *Neuron* **14**, 1303–1309 (1995).
- Wang, K.-W., Tai, K.-K. & Goldstein, S. A. N. MinK residues line a potassium channel pore. *Neuron* **16**, 571–577 (1996).
- Tai, K.-K., Wang, K.-W. & Goldstein, S. A. N. MinK potassium channels are heteromultimeric complexes. *J. Biol. Chem.* **272**, 1654–1658 (1997).
- Romey, G. et al. Molecular mechanism and functional significance of the minK control of the KvLQT1 channel activity. *J. Biol. Chem.* **272**, 16713–16716 (1997).
- Yellen, G., Sodickson, D., Chen, T. Y. & Jurman, M. E. An engineered cysteine in the external mouth of a K⁺ channel allows inactivation to be modulated by metal binding. *Biophys. J.* **66**, 1068–1075 (1994).
- Blumenthal, E. M. & Kaczmarek, L. K. The minK potassium channel exists in functional and nonfunctional forms when expressed in the plasma membrane of *Xenopus* oocytes. *J. Neurosci.* **14**, 3097–3105 (1994).
- Armstrong, C. M. Inactivation of the potassium conductance and related phenomena caused by quaternary ammonium ion injection in squid axons. *J. Gen. Physiol.* **54**, 553–575 (1969).
- Latorre, R. & Miller, C. Conduction and selectivity in potassium channels. *J. Membr. Biol.* **71**, 11–30 (1983).
- Yang, N., George, A. L. Jr & Horn, R. Molecular basis of charge movement in voltage-gated sodium channels. *Neuron* **16**, 113–122 (1996).
- Goldstein, S. A. N. A structural vignette common to voltage sensors and conduction pores: canaliculi. *Neuron* **16**, 717–722 (1996).
- Pascual, J. M., Shieh, C. C., Kirsch, G. E. & Brown, A. M. K⁺ pore structure revealed by reporter cysteines at inner and outer surfaces. *Neuron* **14**, 1055–1063 (1995).
- Dwyer, T. M., Adams, D. J. & Hille, B. The permeability of the endplate channel to organic cations in frog muscle. *J. Gen. Physiol.* **75**, 469–492 (1980).

Acknowledgements. We thank G. Abbott, B. Fakler, P. Ruppersberg and K. W. Wang for helpful discussions. The work was supported by a grant from the NIH-NIGMS to S.A.N.G. We also thank the late Dorothy W. Goldstein, to whom this paper is dedicated.

Correspondence and requests for materials should be addressed to S.A.N.G. (e-mail: steve.goldstein@yale.edu).

Design of potent selective zinc-mediated serine protease inhibitors

Bradley A. Katz*, James M. Clark*, Janet S. Finer-Moore†, Thomas E. Jenkins‡, Charles R. Johnson*, Michael J. Ross*, Christine Luong*, William R. Moore* & Robert M. Stroud*†

* Arris Pharmaceutical Corporation, 385 Oyster Point Boulevard, Suite 3, South San Francisco, California 94080, USA

† Department of Biochemistry, University of California at San Francisco, California 94143, USA

Many serine proteases are targets for therapeutic intervention because they often play key roles in disease¹. Small molecule inhibitors of serine proteases with high affinity are especially interesting as they could be used as scaffolds from which to develop drugs selective for protease targets. One such inhibitor is bis(5-amidino-2-benzimidazolyl)methane (BABIM), standing out as the best inhibitor of trypsin (by a factor of over 100) in a series of over 60 relatively closely related analogues^{2–4}. By probing the structural basis of inhibition, we discovered, using crystallographic methods, a new mode of high-affinity binding in which a Zn²⁺ ion is tetrahedrally coordinated between two chelating nitrogens of BABIM and two active site residues, His57 and Ser195. Zn²⁺, at subphysiological levels, enhances inhibition by over 10³-fold. The distinct Zn²⁺ coordination geometry implies a strong dependence of affinity on substituents. This unique structural paradigm has enabled development of potent, highly selective, Zn²⁺-dependent inhibitors of several therapeutically important serine proteases, using a physiologically ubiquitous metal ion.

At pH 8.2 the inhibition constant (K_i) of BABIM for bovine trypsin in 1.0 mM EDTA, pH 8.2, is 19 μM. Addition of Zn²⁺ at concentrations as low as 100 nM increases the affinity of BABIM for trypsin by 3,800 to K_i = 5.0 ± 0.3 nM. Zn²⁺ alone is a weak inhibitor of trypsin with a K_i of ~33 mM at pH 5.5, and ~1.0 mM at pH 7.0 (concentrations well above physiological levels). A BABIM derivative, bis(5-amidino-2-benzimidazolyl)-methane ketone (keto-BABIM) has an even higher affinity, K_i < 1.0 nM in the presence of Zn²⁺, and a > 1.9 × 10⁴ activity enhancement in the presence of Zn²⁺ versus EDTA.

The Zn²⁺-mediated inhibitor-binding motif, crystallographically determined at pH 8.2 in 1.0 mM Zn²⁺ for trypsin-BABIM-Zn²⁺ and -keto-BABIM-Zn²⁺, is shown in Fig. 1. A peak of 28 ± 1 electrons and temperature factor (B) of 28 Å², corresponds to a Zn²⁺ ion tetrahedrally coordinated by one nitrogen from each of the two benzimidazoles of keto-BABIM, by Ne2_{His57} and by Oγ_{Ser195} (Fig. 1). Bond distances involving Zn²⁺ are similar to those observed in other Zn²⁺-containing proteins^{5,6} and small molecules^{7–9}. The Ser195 hydroxyl may be deprotonated because of coordination by Zn²⁺; the pK_a of water coordinated to Zn²⁺ can be lowered by as much as 9 units^{5–8}.

The entire bound keto-BABIM molecule is well defined by density. As expected, the amidinobenzene portion of BABIM and keto-BABIM superimpose nearly exactly on benzamidine in the structure of trypsin-benzamidine^{10,11}. Two hydrogen-bonded salt bridges are formed, each in the optimal *cis* orientation¹², between amidino amino groups and the carboxylate oxygens of Asp189. A third amidino amino group hydrogen-bonds with Oγ_{Ser190} and with

‡ Present address: Advanced Medicine Inc., 270B Littlefield Avenue, South San Francisco, California 94080, USA.

a bound water, and the fourth with O_{Gly214} (Fig. 2). The distal amidinobenzimidazole of keto-BABIM occupies the S1' site on the surface of the enzyme, its position and orientation tightly coupled to the Zn²⁺ coordination.

The crystal structure of trypsin cocrystallized with keto-BABIM and 1.0 mM Co²⁺ is nearly identical to that of trypsin-keto-BABIM-Zn²⁺. Thus Co²⁺ is also capable of co-inhibition, but assays show that concentrations higher than 25 μ M are required. The structure of trypsin-keto-BABIM in the absence of Zn²⁺ or Co²⁺ shows a direct hydrogen bond between the N3 atom of the S1 benzimidazole and O_{Ser195} showing that high concentrations of Mg²⁺ in the crystallization solution (1.68 M) do not mediate inhibitor binding. The crystal structure of trypsin-BABIM at pH 5.9 in the absence of Zn²⁺ shows that binding of BABIM is mediated by a sulphate ion. The sulphate oxygens form hydrogen bonds with protonated nitrogens from two of the imidazole groups of BABIM, with Ne2_{His57}, O_{Ser195}, N_{Gly193}, and with two water molecules. Addition of 25 mM SO₄²⁻ or PO₄³⁻ improves inhibition by 6- and 10-fold, respectively, at pH 5.9. Thus we have visualized crystallographically three fundamentally different binding mechanisms involving a cation (Zn²⁺ or Co²⁺), an anion (SO₄²⁻) and no ion. These different binding modes reflect the amphoteric nature of the imidazoles of the inhibitors and of His 57, and of the hydroxyl of Ser 195, whose hydrogen-bond donor/acceptor character is dictated by pH and environment.

In the presence of Zn²⁺, the *K_i* of BABIM for trypsin, 5.0 $\pm$ 0.3 nM, corresponds to a free energy of association, ΔG_f^0 , of -11.1 kcal mol⁻¹. Benzamidine alone, corresponding structurally to less than half of BABIM, has a ΔG_f^0 of -6.3 kcal mol⁻¹ (ref. 10). Thus the binding of BABIM to trypsin in the presence of Zn²⁺ is 4.8 kcal mol⁻¹ more favourable than the binding of benzamidine.

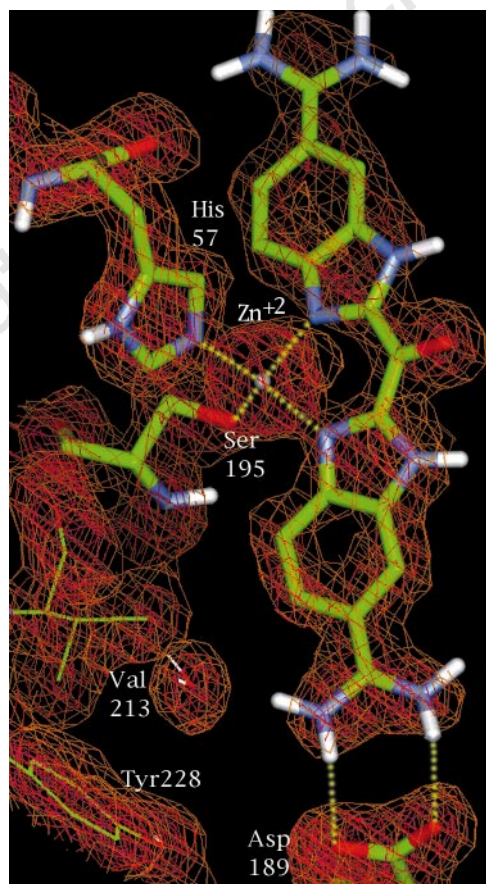


Figure 1 ($2F_o - F_c$), α_c map superimposed on the structure of trypsin-keto-BABIM-Zn²⁺, pH 8.2.

The *K_i* for trypsin of 2-methyl-5-amidinobenzimidazole, corresponding structurally to about half of the BABIM framework, is similar to that of benzamidine (25 μ M)¹⁰, but $\sim 10^3$ -fold weaker than that of BABIM. Thus no additional affinity is gained from the imidazole of 2-methyl-5-amidinobenzimidazole. Without Zn²⁺, BABIM is a weak inhibitor (*K_i* = 19 μ M at pH 8.2). The structure of the amidinobenzimidazole in the S1 pocket is the same in the presence or absence of Zn²⁺. These data demonstrate that concurrent chelation of Zn²⁺ by trypsin and BABIM increases potency by $\sim 10^3$ -fold. The reported *K_i* values^{2,3} for the binding of BABIM to trypsin-like proteases under typical conditions must also reflect the concurrent binding of Zn²⁺, previously unrecognized, because

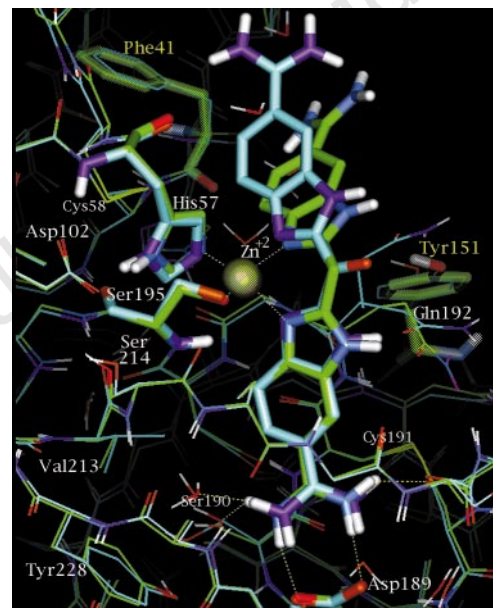


Figure 2 Superposition of trypsin-BABIM-Zn²⁺ onto keto-BABIM-Zn²⁺. In the trypsin-BABIM-Zn²⁺ structure, carbons are green, oxygens red, nitrogens blue and sulphurs yellow. In the trypsin-keto-BABIM-Zn²⁺ structure, carbons are light blue, oxygens orange, nitrogens rose, and sulphurs yellow.

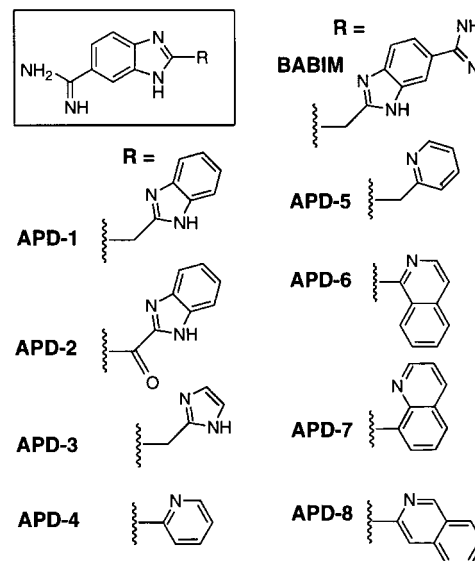


Figure 3 Structures of Zn²⁺-chelating serine protease inhibitor scaffolds.

Table 1 Zinc chelator-based design of potent, selective serine protease inhibitors

Cpd*	Human protease	Inhibition (K_i , μM)†		Zn^{2+} -factor§	Selectivity factor		Selectivity ratio¶
		+ Zn^{2+}	+EDTA‡		+ Zn^{2+}	+EDTA	
BABIM	Trypsin	0.09	18.8	210	90	1.9	47
	Tryptase	0.005	2.5	500	5.0	0.25	20
	Thrombin	0.023	3.7	160	23	0.27	85
	Chymotrypsin	120	710	5.9	120,000	71	1,690
	Chymase	>1,000	>1,000				
	Factor Xa	0.001	10	10,000		Reference enzyme	
APD-1	Trypsin	0.0235	182	7,700	4.2	79	0.05
	Tryptase	0.0695	13.5	190	12	5.8	2.1
	Thrombin	0.0056	2.31	410		Reference enzyme	
APD-2	Trypsin	0.0052	87.5	17,000		Reference enzyme	
	Tryptase	0.05	5.72	110	9.6	0.07	140
	Thrombin	0.101	>1,000	9,900	2.0	>11	<0.18
APD-3	Trypsin	0.666	273	410	29	0.11	260
	Tryptase	1.65	28.8	17	71	1.9	37
	Thrombin	0.0231	15.3	660		Reference enzyme	
APD-4	Trypsin	20.9	101	5.0	29	1.2	2.4
	Tryptase	1.6	19.3	12	2.2	0.23	9.6
	Thrombin	0.721	81.5	110		Reference enzyme	
APD-5	Trypsin	136	>1,000	7.4	440	~1	~440
	Tryptase	0.31	358	1,200		Reference enzyme	
	Thrombin	10.5	>1,000	95	34	>2.8	<12
APD-6	Trypsin	22.5	31.2	1.4	550	1.0	550
	Tryptase	54.5	8.8	0.16	1,330	0.28	4,750
	Thrombin	0.041	31	760		Reference enzyme	
APD-7	Trypsin	8.04	41.8	5.2	25	2.1	12
	Tryptase	11	11	1.0	34	0.55	62
	Thrombin	0.321	20.1	60		Reference enzyme	
APD-8	Trypsin	16.4	56.6	3.5	8.9	2.2	4.0
	Tryptase	63.3	23.4	0.37	34	1.0	34
	Thrombin	1.85	25.5	14.0		Reference enzyme	

* Compound structures are shown in Fig. 3.

† Assay conditions are described in the text.

‡ Values of >1,000 μM (1 mM) indicate that the compound was not assayed at concentrations above that level.

§ Affinity in the presence of Zn^{2+} /affinity in the presence of EDTA.

|| The ratio of K_i for each enzyme versus K_i for the enzyme that shows highest affinity (lowest K_i) in presence of Zn^{2+} .

¶ Ratio of selectivity factor in the presence of Zn^{2+} versus selectivity factor in the presence of EDTA.

exclusion of Zn^{2+} at trace levels where dramatic enhancement in inhibition of BABIM occurs, requires special procedures.

The new Zn^{2+} -mediated mode of protease inhibition described here has therapeutic applications. The catalytic triads of all available crystal structures of active trypsin-like serine proteases are virtually identical and have a similar structural relationship with the S1 pocket. Thus all these proteases, many of which are involved in disease conditions, are capable of coordinating Zn^{2+} in the presence of BABIM-like chelators, and therefore susceptible to potent Zn^{2+} -mediated inhibition. Cysteine proteases could also be inhibited using Zn^{2+} as a co-inhibitor. However, because the relationship of the S1 site and the catalytic His and Cys groups is different from that of the corresponding groups in serine proteases, inhibitor scaffolds distinctly different from that of BABIM or keto-BABIM are required for Zn^{2+} -mediated inhibition of cysteine proteases. Many other enzymes use nucleophilic groups at their active sites capable of coordinating metal ions, and thus it is feasible to harness metals as co-inhibitors of enzymes in general by structure-based design.

Specificity for inhibition of target proteases is conferred both inside and outside the S1 pocket¹³. Substrates of many trypsin-like serine proteases use a Lys or Arg that forms a salt bridge(s) with Asp 189 at the base of the S1 pocket. Most inhibitors of such proteases including BABIM mimic this interaction. Other trypsin-like serine proteases such as chymotrypsin and chymase have an uncharged side chain such as Ser at position 189. Thus the inhibitor group targeted to the S1 pocket of this subclass should also be neutral and structurally complementary. A key requirement of an inhibitor that uses the metal chelation motif described here is that the S1 recognition element appropriate for the particular protease subclass be an appropriate distance from the Zn^{2+} chelating groups

in an orientation that allows simultaneous coordination of the Zn^{2+} by the inhibitor and by the active site His 57 and Ser 195 side chains.

Other determinants of specificity of trypsin-like proteases for compounds based on this motif occur in the regions of residues 34 to 41 and 143 to 154, outside the more conserved P1 site, where there are distinct sequence and structural differences among related proteases¹³. In trypsin, residues Phe41 and Tyr151 from these regions are labelled in yellow in Fig. 2. In the structure of trypsin-BABIM- Zn^{2+} , the distal benzimidazole is 3.4 Å from C β _{Phe41}, 4.2 Å from C ϵ ₂_{Tyr39}, and 6.5 Å from O η _{Tyr151}. In thrombin an additional major determinant of specificity is provided by a loop absent in the other trypsin-like proteases. This loop, inserted between residues 60 and 61, is in the immediate vicinity of the distal benzimidazole of BABIM or keto-BABIM when they are docked into thrombin as metal complexes analogous to those described here, facilitating design of Zn^{2+} -chelating compounds specific for, or against, thrombin. Planar keto-BABIM and non-planar BABIM probe the S1' and S2' sites, respectively (Fig. 2), providing complementary frameworks for designing specificity.

The generality of the Zn^{2+} chelator inhibitor design paradigm is indicated by the structural diversity of scaffolds developed both to function as serine protease inhibitors and to show enhanced inhibition and selectivity in the presence of Zn^{2+} (Fig. 3). Within a series designed to inhibit the trypsin-like serine proteases, trypsin, tryptase and thrombin, varied compound structures (Fig. 3) serve well as scaffolds, exhibiting differential patterns of potency and selectivity (Table 1). Enhancements in potency of compounds in Table 1 in the presence of Zn^{2+} by up to 17,000 (for example, for inhibition of trypsin by APD-2) have been achieved. For each compound, Table 1 also provides the selectivity factor, both in the

Table 2 Crystallography of trypsin-BABIM and -keto-BABIM complexes

	Trypsin-BABIM			Trypsin-keto-BABIM		
	Zn ²⁺	Zn ²⁺	SO ₄ ²⁻	Zn ²⁺	Co ²⁺	Zn ²⁺ -free
Parameters*						
Protein databank accession code	1XUG	1XUF	1XUK	1XUJ	1XUH	1XUI
pH	8.2	8.2	5.9	8.2	8.2	8.2
Data collection system	R-AXIS IV	X-1000	R-AXIS II	X-1000	X-1000	R-AXIS IV
Space group	<i>P</i> 2 ₁ 2 ₁ 2 ₁ †	<i>P</i> 2 ₁ 2 ₁ 2 ₁ ‡	<i>P</i> 2 ₁ 2 ₁ 2 ₁ ‡	<i>P</i> 3 ₁ 2 ₁	<i>P</i> 3 ₁ 2 ₁	<i>P</i> 3 ₁ 2 ₁
No. atoms (including disorder)	4,135§	1,995	1,926	1,834	1,816	1,843
No. waters (including ions)	204	214	211	146	145	157
No. discretely disordered groups	17	21	20	9	18	11
No. side chains with refined occs¶	14	13	16	19	13	24
Diffraction statistics						
No. observations#	97,610	14,115	43,750	55,659	15,007	69,702
Average redundancy	2.6	1.6	3.5	4.1	2.1	2.7
<i>R</i> _{merge} (%)‡‡	8.8	8.5	7.9	8.1	12.8	6.7
Refinement statistics						
Resolution (Å)	7.5–1.50	8.0–2.04	8.0–1.87	7.5–1.92	7.5–2.37	7.5–1.50
No. merged reflections to	33,189	8,337	11,673	12,502	6,599	22,297
<i>F</i> _o /σ cutoff	2.0	1.0	1.0	2.0	1.0	1.8
<i>R</i> _{cryst} (%)**	20.6	15.3	13.0	14.4	13.5	19.8
Free <i>R</i> _{cryst} (%)††	22.6	22.9	17.1	17.1	18.8	22.3
Overall completeness	73.6	57.0	63.7	80.7	76.4	71.9
Completeness at highest resolution	46.8	51.0	33.5	41.2	47.1	32.2
Highest resolution shell (Å)	1.57–1.50	2.13–2.04	1.95–1.87	2.01–1.92	2.47–2.37	1.57–1.50
r.m.s. Deviations‡‡						
Bond lengths (Å)	0.017	0.017	0.015	0.016	0.018	0.018
Bond angles (°)	3.8	3.7	2.9	3.1	3.3	3.3
Torsion angles (°)	26.1	26.5	26.3	26.8	27.0	26.8

* Restrained, isotropic temperature factors were refined and bulk solvent contributions included.

† Large cell.

‡ Small cell.

§ All hydrogens were included. For other structures polar hydrogens were included in the force field during refinement but not in the structure factor or map calculations and are not included in the total atom counts.

|| Not including waters.

¶ Density for all side-chain or terminal atoms in these groups was weak or absent and temperature factors were high. Occupancies for poorly defined groups of atoms were refined. Discretely disordered groups are not included in this category.

Siemens (X-1000) data with *R*_{sym} > 50% were rejected along with data with values >3.5σ from the mean for each set of symmetry equivalents. R-Axis II and IV data were rejected if $\langle (h_i - \langle h \rangle) / \langle (h_i - \langle h \rangle) \rangle \rangle > (0.30 \cdot \langle (global I) \rangle + 0.10 \cdot \langle (h) \rangle)$, where $\langle (h) \rangle$ is the *i*th observation of the intensity of reflection *h*.

‡‡ *R*_{merge} = $\sum_i \sum_j |I(h_i) - \langle I(h) \rangle| / \sum_i \sum_j I(h_i)$.

** *R*_{cryst} = $\sum_i (|F_o| - |F_c|) / \sum_i |F_o|$ (for reflections from 7.5 Å to the highest resolution).

†† Cross validation *R*-factor using 10% of the data withheld from the refinement³⁰.

‡‡ Root mean square deviations from ideal bond lengths, bond angles and torsions.

presence of Zn²⁺ and EDTA. This factor, defined as the ratio of the *K_i* for each enzyme versus the *K_i* for the enzyme that is the most strongly inhibited in the presence of Zn²⁺, quantifies the selectivity of inhibition in the presence or absence of Zn²⁺. The ratio of the selectivity factor in the presence of Zn²⁺ to that in the presence of EDTA (last column of Table 1) shows the effect of Zn²⁺ chelation in enhancing sensitivity to inhibitor substituents outside the S1 and active sites. This selectivity enhancement factor is often quite large, up to 4,750 for inhibition of thrombin versus trypsin by APD-6, and is less than unity only twice out of 20. Thus Zn²⁺ can be harnessed not only to enhance potency but also to amplify selectivity through the requirement of a distinct chelation geometry that is sensitive to the interactions of the distal groups of the inhibitors with regions of the enzymes whose structures differ from one another.

The structural simplicity of the scaffolds shown in Fig. 3 has allowed synthesis of widely varied analogues with substitution patterns that exhibit further enhanced affinity and specificity for targets. Specificity is achieved through capture of favourable interactions in the target versus loss or impairment of such interactions in a non-target, for example by steric clash, or electrostatic incompatibility, or through steric restraints at specificity sites that modulate the Zn²⁺ coordination geometry at the active site. Elaborations in the distal portions of the scaffolds in Fig. 3 from substituents directed to sites that vary among different proteases (see above) have enabled development of potent (*K_i* < 1 nM), selective, Zn²⁺-mediated inhibitors of specific protease targets such as trypsin, thrombin, factor Xa and chymase.

BABIM is relatively non-toxic to cells, or to rats for which the half-maximal lethal dose (LD₅₀) is 130 mg kg⁻¹ (ref. 14) corresponding to ~300 μM. Such levels are ~10⁴- to 10⁵-fold higher than that

required to inhibit trypsin. Zn²⁺ is ubiquitous in human fluids, at levels corresponding to 60 to 120 μM in tissue and blood^{15,16}, although the concentrations of free Zn²⁺ are significantly lower because of chelation by metalloproteins. Nevertheless, metallo-inhibitors that bind Zn²⁺ strongly enough to compete for and harness Zn²⁺ from physiological fluids and use it for target recognition under physiological conditions are being developed as leads to efficacious pharmaceuticals for treating diseases associated with a wide spectrum of serine proteases.

Complexes with transition metals (mostly non-physiological) have previously been used to inhibit proteins^{17,18}. Zn²⁺ itself inhibits some enzymes such as cytomegalovirus protease¹⁹, cysteine proteases and HIV protease⁵ by binding at their active sites. The crystal structure of an enzyme target complexed with Zn²⁺ alone at the active site provides a strong starting point from which to design potent small molecule inhibitors that use Zn²⁺ as an integral component of binding. Protein engineering studies have also sought to harness Zn²⁺ for mediating macromolecular recognition through modification of one of the counterparts of protein–protein or protein–peptide complexes by site-directed mutagenesis^{20–22}. Our discovery, that physiological metal ions can be recruited for mediating binding of small molecules to unmodified natural targets, opens promising new discovery routes for structure-based drug design. □

Methods

Synthesis and characterization of BABIM and keto-BABIM. BABIM was synthesized as described previously^{2,23}. Spontaneous conversion of clear BABIM to yellow keto-BABIM solution was effected by air oxidation in DMSO and monitored by high-performance liquid chromatography (HPLC)

and nuclear magnetic resonance (NMR). Keto-BABIM was purified to a single component by HPLC. NMR and mass spectrometry were used to establish its chemical structure.

Enzymes and inhibition assays. Human and bovine trypsin and human chymotrypsin were from Athens Research Technology. Human thrombin and factor Xa were from Calbiochem and Pharmacia & Upjohn, respectively. Human chymase²⁴ and trypsin²⁵ were produced at Arris. K_i s of compounds for trypsin, trypsin, thrombin and factor Xa were determined by monitoring hydrolysis at 410 nm of the substrate tosyl-Gly-Pro-Lys-p-nitroanilide (Sigma), by 1.0 nM enzyme in 50.0 mM Tris, 100 or 1,000 mM NaCl, 10% DMSO, pH 8.2 at 22 °C in 150 mM Zn²⁺ or 10 mM EDTA. For bovine trypsin the effect of Zn²⁺ and other ions at various concentrations was also assayed in 1.0 mM Tris, pH 8.2 or 1.0 mM MES, pH 5.5. For factor Xa the substrate was MeOC-Nle-Gly-Arg-pNa, and for chymase and chymotrypsin it was Suc-Val-Pro-Phe-SBzl. All buffers contained 0.050% Tween-20. Enzymes were incubated for 1 h with inhibitor at eight inhibitor concentrations bracketing the K_i , prepared by serial dilution along with eight controls lacking the inhibitor. Substrate was added, and initial rates of substrate hydrolysis were determined using a UV/MAX Kinetic Microplate Reader (Molecular Devices). K_i (apparent) values were determined with BatchKi or true K_i values with DYNAFIT²⁶.

Crystallography. Crystals of bovine trypsin-benzamidine were grown at pH 8.2 in the small ($a = 54.8 \text{ \AA}$, $b = 58.7 \text{ \AA}$, $c = 67.6 \text{ \AA}$) or large ($a = 63.7 \text{ \AA}$, $b = 63.5 \text{ \AA}$, $c = 69.0 \text{ \AA}$) orthorhombic cells. Crystals of trypsin-BABIM, pH 8.2, were prepared by soaking trypsin-benzamidine crystals in synthetic mother liquor (1.62 M MgSO₄·7 H₂O, 100 mM Tris saturated with BABIM, and containing 1.0 mM or 5.0 mM ZnSO₄). The soaking solutions were replaced once a day over a period of 4 days with freshly prepared solutions. Trypsin-keto-BABIM-Zn²⁺ was cocrystallized in space group $P3_121$ from solutions saturated in keto-BABIM and 1.0 mM in Zn²⁺. Crystals of trypsin-BABIM-SO₄²⁻, pH 5.9, were prepared in the absence of Zn²⁺ by soaking trypsin-benzamidine crystals in synthetic mother liquor buffered with 100 mM MES and saturated with BABIM.

Structures were determined by described procedures^{11,27}. X-ray diffraction data were collected on a multiwire area IPC X-1000 detector from Siemens, or on an R-AXIS II or an R-AXIS IV image plate detector from Molecular Structure Corporation (MSC). Data were processed with the XDS²⁸, or with Biotex from MSC. Difference Fourier maps with coefficients ($|F_o| - |F_c|$), or ($2|F_o| - |F_c|$), and phases α_o , where $|F_o|$ and $|F_c|$ denote observed and calculated structure factors amplitudes, respectively, yielded initial structures which were refined to convergence with X-PLOR²⁹. Data collection and refinement statistics are given in Table 2.

Received 1 September; accepted 27 October 1997.

1. Stroud, R. M. A family of protein-cutting proteins. *Sci. Am.* **231**, 74–88 (1974).
2. Tidwell, R. R. & Geratz, J. D. Diarylamidine derivatives with one or both of the aryl moieties consisting of an indole or indole-like ring. Inhibitors of arginine-specific esterproteinases. *J. Med. Chem.* **21**, 613–623 (1978).
3. Jennette, J. C., Tidwell, R. R., Geratz, J. D., Bing, D. H. & Falk, R. J. Amelioration of immune complex-mediated glomerulonephritis by synthetic protease inhibitors. *Am. J. Pathol.* **127**, 499–506 (1987).

4. Caughey, G. H., Raymond, W. W., Bacci, E., Lombardy, R. J. & Tidwell, R. R. Bis(5-amidino-2-benzimidazolyl)methane and related amidines are potent, reversible inhibitors of mast cell tryptases. *J. Pharmacol. Exp. Ther.* **264**, 676–682 (1993).
5. Lipscomb, W. N. & Sträter, N. Recent advances in zinc enzymology. *Chem. Rev.* **96**, 2375–2433 (1996).
6. Christianson, D. W. Structural biology of zinc. *Adv. Protein Chem.* **42**, 281–355 (1991).
7. Kimura, E., Shiota, T., Shiro, M. & Kodama, M. A zinc(II) complex of 1,5,9-triazacyclodecane ([12]aneN₃) as a model for carbonic anhydrase. *J. Am. Chem. Soc.* **112**, 5805–5811 (1990).
8. Groves, J. T. & Olson, J. R. Models of zinc-containing proteases. Rapid amide hydrolysis by an unusually acidic Zn²⁺-OH₂ complex. *Inorg. Chem.* **24**, 2715–2717 (1985).
9. Vedani, A. & Huhta, D. A new force field for modeling metalloproteins. *J. Am. Chem. Soc.* **112**, 4759–4767 (1990).
10. Krieger, M., Kay, L. M. & Stroud, R. M. The structure and specific binding of trypsin: A comparison of inhibited derivatives and a model for substrate binding. *J. Mol. Biol.* **83**, 209–230 (1974).
11. Katz, B. A., Finer-Moore, J. S., Mortezaei, R., Rich, D. H. & Stroud, R. M. Novel K_i ~ nanomolar inhibitors of serine proteases by binding or epitaxial chemistry on an enzyme surface. *Biochemistry* **34**, 8264–8280 (1995).
12. Ippolito, J. A., Alexander, R. S. & Christianson, D. W. Hydrogen bond stereochemistry in protein structure and function. *J. Mol. Biol.* **215**, 457–471 (1990).
13. Perona, J. J. & Craik, C. S. Structural basis of substrate specificity in the serine proteases. *Protein Sci.* **4**, 337–360 (1995).
14. Tidwell, R. R., Geratz, J. D., Clyde, W. A., Rosenthal, K. U. & Dubovi, E. J. Suppression of respiratory syncytial virus infection in cotton rats by bis(5-amidino-2-benzimidazolyl)methane. *Antimicrob. Agents Chemother.* **26**, 591–593 (1984).
15. Lenter, C. (ed.) *Geigy Scientific Tables Vol. 3 Physical Chemistry Composition of Blood Hematology Somatometric Data* 87 (Ciba-Geigy, Basle, Switzerland, 1984).
16. Hambridge, K. M., Casey, C. E. & Krebs, N. F. in *Trace Elements in Human and Animal Nutrition* 5th edn vol. 2 (ed. Mertz, W.) 15 (Academic, Orlando, Florida, 1996).
17. Lippard, S. J. in *Bioinorganic Chemistry* (eds Bertini, I., Gray, H. B., Lippard, S. J. & Valentine, J. S.) 505–583 (University Science, Mill Valley, CA, 1994).
18. Berners-Price, S. J. & Sadler, P. J. Coordination chemistry of metalodrugs: insights into biological speciation from NMR spectroscopy. *Coord. Chem. Rev.* **151**, 1–40 (1996).
19. Tong, L. et al. A new serine-protease fold revealed by the crystal structure of human cytomegalovirus protease. *Nature* **383**, 272–275 (1996).
20. Matthews, D. J. & Wells, J. A. Engineering an interfacial zinc site to increase hormone-receptor affinity. *Chem. Biol.* **1**, 25–30 (1994).
21. Somers, W., Ullsch, M., De Vos, A. M. & Kossiakoff, A. A. The X-ray structure of a growth hormone-prolactin receptor complex. *Nature* **372**, 478–481 (1994).
22. Brinen, L. S., Willett, W. S., Craik, C. S. & Fletterick, R. J. X-ray structures of a designed binding site in trypsin show metal-dependent geometry. *Biochemistry* **35**, 5999–6009 (1996).
23. Tidwell, R. R., Geratz, J. D. & Dubovi, E. J. Aromatic amidines: comparison of their ability to block respiratory syncytial virus induced fusion and to inhibit plasmin, urokinase, thrombin, and trypsin. *J. Med. Chem.* **26**, 294–298 (1983).
24. McGrath, M. E. et al. Production of crystallizable human chymase from a *Bacillus subtilis* system. *FEBS Lett.* **413**, 486–488 (1997).
25. Tanaka, R. D., Clark, J. M., Abraham, W. M. & Moore, W. R. Mast cell tryptase: a new target for therapeutic intervention in asthma. *Int. Arch. Allergy Immunol.* **107**, 408–409 (1995).
26. Kuzmic, P. Program DYNAFIT for the analysis of enzyme kinetic data: application to HIV proteinase. *Anal. Biochem.* **237**, 260–273 (1996).
27. Katz, B. A. & Cass, R. T. In crystals of complexes of streptavidin with peptide ligands containing the HPQ sequence the pK_a of the peptide histidine is less than 3.0. *J. Biol. Chem.* **272**, 13220–13228 (1997).
28. Kabsch, W. Evaluation of single-crystal X-ray diffraction data from a position-sensitive detector. *J. Appl. Crystallogr.* **21**, 916–924 (1988).
29. Brünger, A. T. in *X-PLOR Manual, version 3.1: A System for X-ray Crystallography and NMR* 187–206 (Yale University, New Haven, CT, 1990).
30. Brünger, A. T. The free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472–474 (1992).

Acknowledgements. We thank K. Spear for synthesis of BABIM, T. Church and M. Linsell for synthesis of Arris compounds, L. Cregar, D. Putnam and R. Warne for enzyme assays, and M. Venuti for review and discussion of the manuscript.

Correspondence and requests for materials should be addressed to B.A.K. (e-mail: bak@arris.com). Brookhaven Databank accession codes are given in Table 2.

KNOW YOUR COPY RIGHTS RESPECT OURS

The publication you are reading is protected by copyright law. Photocopying copyright material without permission is no different from stealing a magazine from a newsagent, only it doesn't seem like theft.

If you take photocopies from books, magazines and periodicals at work your employer should be licensed with CLA.

Make sure you are protected by a photocopying licence.



The Copyright Licensing Agency Limited
90 Tottenham Court Road, London W1P 0LP
Telephone: 0171 436 5931 Fax: 0171 436 3986

Telemicroscopy via the Internet

We report a concept that is not only applicable for telemicroscopy, but may be adapted for the remote control of any kind of scientific device, opening a new field of application for the Internet.

**Guenter Wolf, Detlev Petersen,
Manfred Dietel and Iver Petersen**

The introduction of Java technology will greatly extend the functionality of the Internet as it allows users to run programs from within a Web page. Until now, this has mainly been used to create visual animated and interactive Web pages. We used it to control a remote device, in this case a telemicroscope.

Server and client

The telemicroscopy system consists of two elements: the microscope-server and the telemicroscopy clients. A computer on the microscope side of the network controls an automatic microscope (Leica's DMRXA) mounted with a CCD camera and function as an Internet server. The microscope-server receives microscope-operation and image-capture commands, executes these commands, converts the captured images into compressed JPEG image files and distributes these image files to the connected telemicroscopy clients.

For the design of the telemicroscopy client we used the Java technology. Java enables the development of programs that are independent of the hardware and operating system. For example, a program written in Java can run on a Windows PC, a Macintosh computer or a UNIX workstation without any system-dependent modifications.

Even better, Internet browsers, such as Netscape or Microsoft Internet Explorer, can execute so-called Java Applets, which are Java programs that can be incorporated into a browser. Our telemicroscopy client program is such a Java Applet and is automatically downloaded and started when a user selects the Web site of the corresponding microscope-server with an Internet browser.

Remote control

The user can move the microscope stage, change the magnification or execute any other microscope operations by pressing the related buttons of the downloaded telemicroscopy client program (Fig. 2). The microscope stage can be positioned relative to the current position by pressing the corresponding direction buttons or by clicking the intended position in a previously captured image of lower magnification. In the image-discussion mode, a user can point to a region of interest on the image by pressing a mouse button. An arrow pointing to this region will be

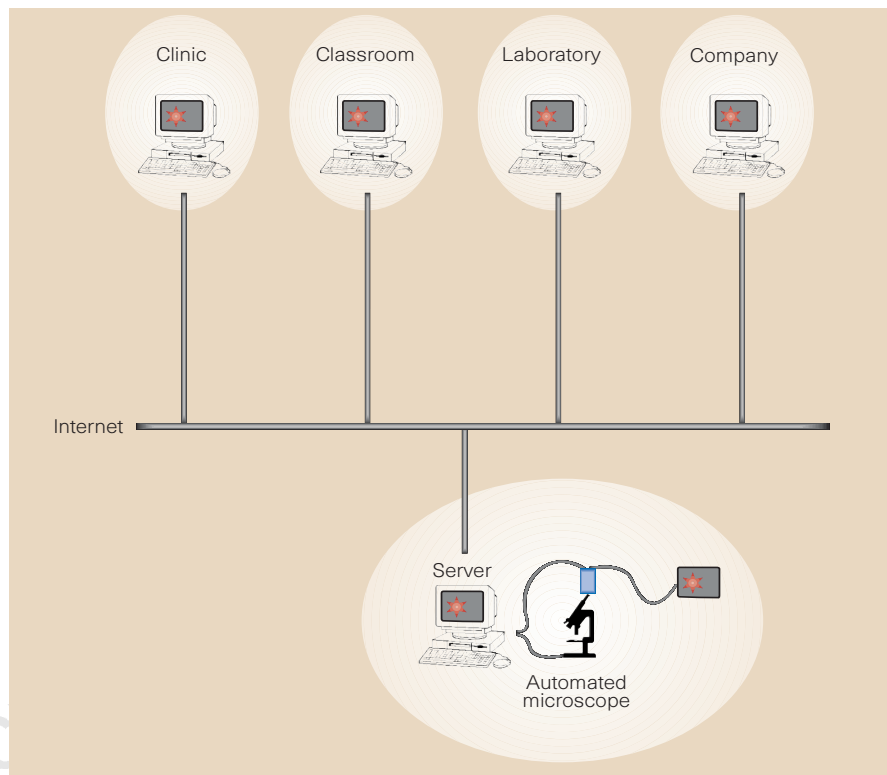


Figure 1 From clinic to classroom: the Internet can link many computer users to the same microscope.

shown to all connected telemicroscopy clients simultaneously. Thus, any Internet user can control the remote microscope and discuss the images with other users that have direct access to the Internet or are connected by modem with ordinary phone lines or through the Integrated Services Digital Network (ISDN). No specific hardware or software is required for the client. Any Internet browser with Java support makes an Internet user a telemicroscopy client. This system can be tested at <http://amba.rz.charite.hu-berlin.de/telemic> (on-line instructions are provided).

The discussion of images and consultation of specialists are an important application of telemicroscopy. In contrast to most of the currently available telemicroscopy systems, no special hardware or software is necessary for the telemicroscopy clients. This is the great advantage of our Internet browser-based solution, which makes every Internet user into a potential consultant. We think that this will help to overcome one essential obstacle to the practicable application for consultations, which is currently the limited number of compatible telemicroscopy clients.

There exist meanwhile further telemicroscopy systems using the Internet as a

telecommunication link or allowing remote control by an Internet browser^{1,2}, but we are not aware of any that uses Java technology for interactive communication in telemicroscopy.

The Java advantage

The use of Java has several advantages compared to static Web pages. This is mainly due to the fact that it enables the efficient distribution of tasks between the server and the clients, thereby minimizing the need for Internet communication. For example, it allows the synchronization of the image display to the clients, whereby the server informs all connected clients about changes to the microscope and the new images appear automatically on the client's browser. Another aspect is the possibility for interactive communication between clients, which is illustrated by the facility of making a pointer visible to all attached clients in the image-discussion mode. These features would be impossible using static Web pages.

The utility of telemicroscopy is clearly illustrated when it is applied to pathology. In other fields of telemedicine, such as radiology and cytology, it is sufficient to inspect a collection of still images, which can be sent by e-mail. The evaluation of histologi-

cal slides, however, requires direct connection to a microscope, allowing the user to find the suspect regions and to get an idea of the three-dimensional structure of the investigated tissue by inspecting a two-dimensional section. This can only be accomplished by moving the slide and changing the magnification of the microscope.

Response time and image quality

Image quality and response times are critical parameters for the practical use of the system. Small images (360×270 pixels with JPEG compression quality 35) ensure short response times, when the user wants to get an overview of the slide by changing the field of view rapidly; whereas, full-sized images (720×540 pixels with JPEG compression quality 75) can be captured for the inspection of details. The factors that influence the response time are mechanical microscope operations, image capture and compression, and finally image transfer and decompression. Interestingly, image compression was the main determinant out of these factors. The time for image compression depends heavily on the computer performance of the microscope-server. We were able to reduce the response times to 2.5–4 s using a 200 MHz MMX PC with 64 MB of RAM in a local area network. These times did not differ significantly when the microscope was operated through a direct Internet connection from other sites within Germany and Europe, indicating that the Internet is generally not prohibitive as a communication link in telemicroscopy.

Many Internet users do not have direct access and are connected either by modem and ordinary phone line or via ISDN. Both cause an additional delay for image transfer. However, the ISDN connection is preferable to conventional phone lines, as it introduced a delay of only 1–2 s for small images and 6–10 s for full-size images, which is still suitable for practical use.

The image quality depends mainly on the camera, the frame grabber and the image compression. We used a standard CCD video camera with PAL video norm (JVC TK-C1380) and a frame grabber (Matrox Meteor) with a resolution of 768×576 pixels, which is sufficient for the majority of applications in telemicroscopy. Similarly, the image-compression factors were adjusted according to the criteria that have been tested in other systems. There are also digital cameras available that allow image resolutions of $1,600 \times 1,200$ pixels and higher, but using these higher image resolutions with the current computer hardware and Internet technology does increase the response times considerably.

Future directions

The described telemicroscopy system is still in a test phase and several develop-

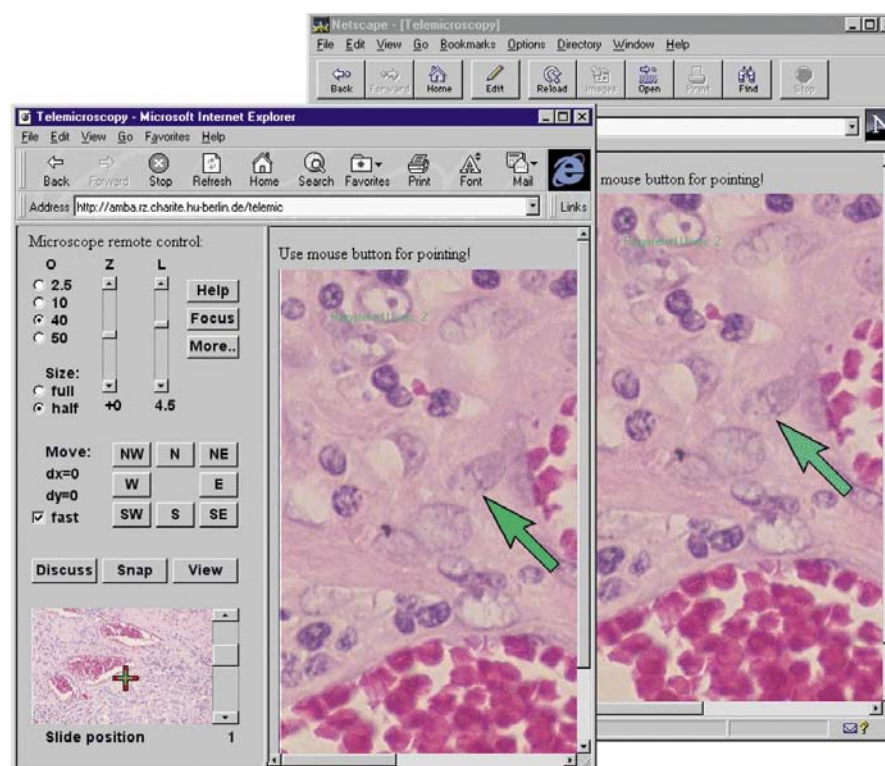


Figure 2 Web pages of two telemicroscopy clients. The microscope can be controlled by pressing the related buttons. The images are transferred simultaneously to all connected Internet users.

ments are envisioned. Most importantly, the performance and the reliability of the Java clients has to be improved as specific browsers in combination with some computer systems cause problems in operating the microscope. Although we found solutions for some severe bugs in the Java Virtual Machines of Microsoft Internet Explorer 4.0 and Netscape Navigator 4.0 for Windows NT and Windows 95 (amba.rz.charite.hu-berlin.de/telemic/feedback/bugs.html), new problems may arise and we would appreciate notification of these by the users. We are confident that differences in the Java implementations between different browser/operating system platforms will be removed in the near future. A number of additional features has been suggested, including the capability to draw and overlay text in the discussion mode, the inclusion of a low-resolution image of the entire slide to allow near real-time operation with lower image resolution and/or a small field of view. Servers for other automatic microscopes are in preparation (Zeiss Axioplan2) or planned. Also requiring support will be new generations of video frame grabbers and digital cameras that operate without a frame grabber — these are currently coming on the market.

Using a conventional microscope

The system can be adapted for the use of conventional microscopes with a CCD camera and monitors for image display and discussion. By recognizing manual changes to

the field of view, the server will transmit the new image for display simultaneously on computer monitors in remote locations. Such a system with a computer, frame grabber and video camera is affordable for any Internet user who wants to discuss their microscope images. The hardware price for such a simple solution, including the frame grabber and video camera, should be less than \$2,000. In contrast, the cost of an automatic microscope could be more than \$50,000. A free test version of the telemicroscopy server software that can be used without an automatic microscope is available from the authors. However, this solution does not enable the remote operation of the microscope.

Our concept is not only applicable for telemicroscopy. It can be adapted for the remote control of any kind of scientific device, opening a new field of application for the Internet. □

Note added in proof: Reference 1 was published during review of this manuscript.

Guenter Wolf, Detlev Petersen*, Manfred Dietel and Iver Petersen: Institute of Pathology, University Hospital Charité, Schumannstr. 20–21, D-10098 Berlin, Germany.

URL: <http://amba.rz.charite.hu-berlin.de/telemic/>
e-mail: wolf@digital-u.com

*DP is with Blood Bank — Immunohematology — Transfusion Medicine, Hannover Medical School, D-30623 Hannover, Germany.

1. Maturro, R., Kath, G., Zeigler, R. & Meehan, P. Control of a Remote Microscope Over the Internet. *Biotechniques* 22, 1154–1157 (1997).
2. <http://www.leica.co.uk/microscope>

Tools for riding the biotechnology wave, including those for designing, synthesizing, analysing and predicting the efficacy of new compounds.

HTS 7000 bioassay reader

From Perkin-Elmer

Designed to be a versatile microplate reader system, which can analyse virtually all biochemical assays, including fluorescence or absorbance (ultraviolet or visible)

This system is appropriate for routine, high-throughput screening of small-volume samples required for drug development applications. Key applications include drug screening and therapeutic drug monitoring, cell-function studies, immunoassays and DNA and protein quantification. The reader can read any plate type, in formats from 6 to 384 wells, using fluorescence, phosphorescence, chemi- or bioluminescence with software that allows the selection of multiple measurement modes.

Reader Enquiry No. 101

Drug metabolizing enzymes

From PanVera

A line of recombinant human phase I and II drug metabolizing enzymes (DMEs)

These highly active recombinant enzymes will aid in the development of new *in vitro* methods for isozyme identification, determination of kinetic parameters for substrates and inhibitors, isolation and synthesis of rare and novel metabolites, and high-throughput screening. PanVera has developed methods for high-level expression of active DMEs, allowing the production of large amounts of recombinant human enzymes in three different formats: Baculosomes (insect microsomes containing an overexpressed recombinant P450 or UGT isozyme), the Recot system (purified P450 isozymes reconstituted in a 'mix-and-metabolize' format) and purified enzymes, including P450 isozymes, GSTs and sulphotransferases.

Reader Enquiry No. 102

Reacto-Stations

From Stem

A series of instruments that allows chemists to easily and inexpensively synthesize compounds of interest

Applications of this instrument include the rapidly expanding fields of combinatorial chemistry, organic synthesis, process optimization, enzyme assays, sample concentration and general incubations. Precise control is provided and, depending on the model, the temperature can be set accurately in the range from -30°C and up to 150°C with the sample simultaneously stirred, either gently or at up to 2,000 r.p.m. These stations

vary in capacity from as small as 10 test tubes up to 50 tubes. The stations operate manually for simple experiments or can be integrated within robotic workstations. The whole device can be controlled remotely through RS 232/485 interfaces, giving the additional facilities of unattended operation, speed variation and temperature ramping.

Reader Enquiry No. 103

SpectraFluoRPlus

From Tecan

A multi-functional microplate reader

Tecan has created this microplate reader to give users a highly flexible and a cost-effective system for applications, such as cell growth and proliferation, ion studies, cell adhesion, cell viability, and others. The system is said to be well suited for laboratories that need a multi-functional system that measures fluorescence or absorbance (in ultraviolet and standard wavelength ranges, as well as performing luminescence and time-resolved measurement). As the instrument performs the functions of several instruments, users benefit by saving valuable time and laboratory space. It reads from 6 to 1,536 wells, as well as Terasaki plates or Petri dishes. A multi-label detection system on the instrument is used to save time, as it detects different fluorescent probes simultaneously on the same microplate.

Reader Enquiry No. 104

Cleanrooms

From Clestra

Comprehensive units that can be used for on-site tissue preparation

Those who process tissues must follow stringent procedures, which this system is designed to accommodate. Entry into the scrub room and corridor takes place through interlocked doors before occupants gain access to the biosafety room. Magnehelic gauges are mounted in each of the rooms to monitor room differentials. The cleanroom suite includes a tissue preparation room, inner and outer equipment rooms, and a stainless steel double-door autoclave, which opens through the stainless steel, double-walled modular partition system. The company's AdvancAir units are used to control the environment in the suite and the ceiling is constructed of the company's load-bearing 5-cm t-grid. In the seven-room project of class-100,000 to class-10,000 areas there are numerous returns and validatable



Cleanrooms for tissue processing.

directional airflow.

Reader Enquiry No. 105

ArgoCaps

From Argonaut Technologies

Pre-weighed resin capsules for solid-phase synthesis

These resin packages offer a convenient and time-saving method for delivery of fixed doses of resin to a reaction vessel. ArgoCaps should eliminate the need for weighing, and protect the resin from moisture. Handling problems associated with sticky or static resins are also eliminated. The capsules are made of a polymer, which is easily dissolved in a wide range of organic solvents. These capsules are currently available in quantities of 50, 100, 500 or 1,000 capsules filled with ArgoPorer, a new family of solid supports based on highly cross-linked macroporous polystyrene and ArgoGel linker products based on a polyethylene glycol-polystyrene graft copolymer. Argonaut also offers custom resin filling of ArgoCaps (minimum quantity per order is 1,000 capsules).

Reader Enquiry No. 106

High-throughput rug discovery

Visage HDG analyser

From Bio Image

A high-density grid (HDG) analyser that is designed to identify and quantify over 200,000 spots in numerous high-density grid images

With this analyser, positives or negatives can be exported to the picking robot to facilitate follow-up experiments. It accommodates all rectangular grid layouts and makes it straightforward to create the reference matrix. A constellation mapping algorithm requires only three anchor points on a reference image in order to overlay multiple study images accurately. This method, unlike rigid-grid methods, compensates for any distortion or skewing that may occur within a sample array. Using proprietary algorithms, the software quantifies the actual spot, not only grid locations. After quantification, clones can be sorted based on the amount of hybridization. Moreover, Visage uses spot-database queries, which can filter information based on the existence of spots, quantitative ratios of matched spots, spot integrated intensities area and user-defined spot characteristics. The HDG analyser supports direct image acquisition, but can also import TIFF format images files from various sources.

Reader Enquiry No. 107

Criterion

From LJI BioSystems

The company's first high-throughput screening (HTS) system

This system consists of Analyst, an auto-



LJL BioSystems: makes light work of assays.

mated, multi-mode assay-detection instrument, which has been designed to connect seamlessly to existing robotics systems; a complete spectrum of application-specific reagents and assay kits, assay development services and application support services. The cornerstone of the system is Analyst, which is capable of performing all four major types of optical detection assays currently in use (for example, fluorescence, fluorescence polarization, time-resolved fluorescence and luminescence assays). According to the manufacturer, it is the only bioassay instrument for accelerated drug discovery that reads assays in either 96- or 384-well microtitre plate formats with no loss of sensitivity or dynamic range. Analyst achieves its improved performance with SmartOptics, a feature that automatically configures the assay detection optics to achieve maximum sensitivity and low background. It is also programmed for plug-and-control and data communications capabilities.

Reader Enquiry No. 108

BioPick

From BioRobotics

A benchtop robot that speeds colony picking and achieves unattended high throughput

According to the manufacturer, over 9,000 colonies can now be picked in one unattended run using a new high-throughput benchtop robot. The compact construction means that it takes up minimum laboratory space. Moreover, the risk of contamination is greatly reduced through the use of on-line plate filling and lid removal. Also, BioPick's needle is sterilized for each use in a constantly replenished ethanol bath. Running off a standard mains supply and requiring no compressed air, the instrument can be installed in any laboratory. It has been designed for reliability, with no flexing wires or moving electrical compo-



The proper way to pick — BioRobotics' BioPick.

nents. Vibration has also been minimized to give quieter operation and reduce air movement and the risk of contamination. As this system is designed to work alongside a gridding instrument, such as the BioGrid, both operations can be carried out simultaneously.

Reader Enquiry No. 109

LOPAC

From RBI

A library of pharmacologically active compounds (LOPAC) for high-throughput screening

The library includes more than 600 pharmaceutically diverse compounds with known drug or drug-like activity. These are well-characterized, high-purity compounds, which can be used to characterize orphan receptors, validate new HTS assays and help to define pharmaceutically relevant screening strategies. The compounds have been organized according to pharmacological class, including adenosines/purinergics; adrenergics/histaminergics; cholinergics/ion channel modulators; dopaminergics; serotonergics; glutamatergics, and others. Each group is organized in HTS-compatible, 96-well racks containing 2 mg of each compound per tube. The eight-rack set is accompanied by plate maps outlining the position of each compound, its chemical name and its corresponding RBI catalogue number. Each compound is also represented in SD format in ISIS base electronic files, which are included. Scale-up and custom services are also available.

Reader Enquiry No. 110

Haystack

From The Automation Partnership

The system is designed to manage and archive a store of all samples, made or acquired, in the drug discovery process

Haystack is designed to manage growing compound collections more effectively and to help build a set of compounds in a liquid store, speeding the overall progress of drug discovery. The automated library management system makes it possible to utilize the vast compound collections that lie at the heart of high-throughput screening. Haystack can accommodate compound libraries running to millions of samples, allowing almost limitless scale-up. Within the system, individual samples are identified by bar codes and compounds are stored in carousels with robotic retrieval systems and conveyors delivering samples to a dispensing area. Haystack can be configured to hold a variety of libraries in different phases ranging from solid archives to liquid libraries in micro-tubes or micro-plates. Its modular design allows it to be readily adapted to differing needs.

Reader Enquiry No. 111

Liquid handling

BioScript Pro

From Beckman Instruments

Programming language for the Biomek 2000 laboratory automation workstation

This upgrade allows researchers with special needs to develop customized programs and to perform complex functions that go beyond the basic capabilities of BioWorks software with standard BioScript. An eight-page bulletin, T-1843A, describes situations in which this programming is particularly useful. BioScript Pro allows users to programme such tasks as pipetting into and out of unusual labware and pipetting with custom parameters that cannot be set in the standard software. This tool is also useful for programming methods that can obtain and manipulate data in DDE server applications, such as Excel, Access, Word and customer-created applications.

Reader Enquiry No. 112

Hydra microdispenser

From Robbins Scientific

A new 12-page, four-colour brochure covers this complete line of microdispensers

According to the manufacturer, the Hydra-384 is the first instrument to provide microlitre dispensing without the use of pipettes tips. It also contains performance information regarding dispensing precision and accuracy for all models. These microdispensers are said to be widely used in genome mapping and sequencing operations, dispensing template DNAs and reagents used in PCR and cycle sequencing reactions. They are also useful for assays used in high-throughput screening of drug libraries.

Reader Enquiry No. 113

These notes are compiled by Brendan Horton from information provided by the manufacturers.

For more details, fill in the reader service card bound inside the journal.

ADVERTISEMENT

NEWCASTLE UNIVERSITY
MOLECULAR BIOLOGY UNIT



PROTEIN SEQUENCING

Membrane-bound, solid or solution samples

£12 per cycle

5 cycles minimum - no set up charge

For further information contact Dr. Joe Gray

Phone: +44 191 222 8612

Fax: +44 191 222 8129

Email: JOE.GRAY@newcastle.ac.uk

READER ENQUIRY NO. 14